

CardioBench: Do Echocardiography Foundation Models Generalize Beyond the Lab?

Darya Taratynova¹, Ahmed Aly¹, Numan Saeed¹, Mohammad Yaqub¹

¹ Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
 {darya.taratynova, ahmed.aly, numan.saeed, mohammad.yaqub}@mbzuai.ac.ae

Abstract

Foundation models are reshaping medical imaging, yet their application in echocardiography remains limited, hindered by a heavy reliance on private datasets that prevent reproducible comparison. Echocardiography poses unique challenges, including noisy acquisitions, high frame redundancy and limited diverse public datasets. To address this, we introduce CardioBench¹², a comprehensive benchmark for echocardiography foundation models. Specifically, CardioBench unifies eight publicly available datasets into a standardized suite spanning four regression and five classification tasks, covering functional, structural, diagnostic, and view recognition endpoints. Leveraging this framework, we evaluate several leading foundation models, including cardiac-specific, biomedical, and general-purpose encoders, under consistent zero-shot, probing, and alignment protocols. Our analysis reveals that while general-purpose encoders transfer well and often close the gap with probing, they struggle significantly with fine-grained distinctions like view classification and subtle pathology recognition. Results indicate that models capturing temporal cardiac dynamics perform best on functional tasks, while retrieval-based approaches generalize more consistently across datasets. By releasing preprocessing, splits, and public evaluation pipelines, CardioBench establishes a reproducible reference point to guide the architectural design of future echocardiography and possibly other medical imaging foundation models.

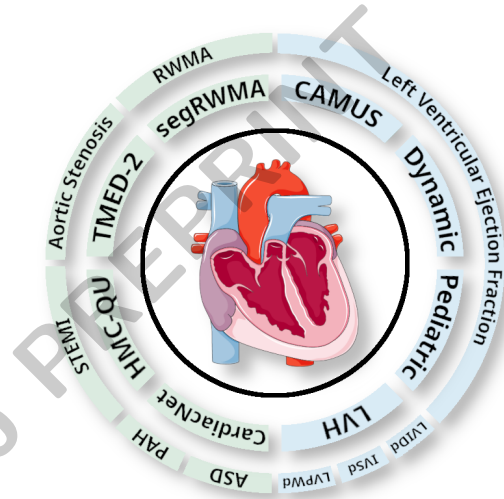


Figure 1: CardioBench: A unified benchmark for echocardiography.

1 Introduction

Foundation models have become a transformative force in vision and language domains, demonstrating remarkable zero-shot generalization across diverse tasks including image classification, retrieval, visual grounding, and multimodal reasoning (Jia *et al.*, 2021; Ghiasi *et al.*, 2022; Li *et al.*, 2022; Singh

et al., 2022; Alayrac *et al.*, 2022]). Large-scale architectures such as CLIP, DINOv3, and SigLIP2 demonstrate that self-supervised and multimodal learning produce general-purpose backbones with strong transferability across downstream tasks ([Radford *et al.*, 2021; Siméoni *et al.*, 2025; Tschannen *et al.*, 2025]). This success has extended to medical imaging, where foundation models have advanced disease classification in chest radiography ([Irvin *et al.*, 2019; Johnson *et al.*, 2019]) and achieved state-of-the-art segmentation in volumetric CT and MRI ([Roy *et al.*, 2023; Huang *et al.*, 2023]). These advances have been enabled by the availability of large, standardized public datasets that allow for reproducible benchmarking and fair model comparison.

In contrast, echocardiography remains underserved despite its central role in clinical practice. It is a first-line, non-invasive imaging modality for assessing cardiac structure and function, and is integral to routine cardiovascular diagnosis and management ([Mitchell *et al.*, 2019]). It is used to evaluate a wide range of cardiac conditions from multiple views, necessitating models that can generalize across overlapping tasks rather than addressing each in isolation. Therefore, there is growing interest in developing ultrasound foundation

¹Code: <https://github.com/BioMedIA-MBZUAI/CardioBench>²Supplementary Material: <https://arxiv.org/abs/2510.00520>

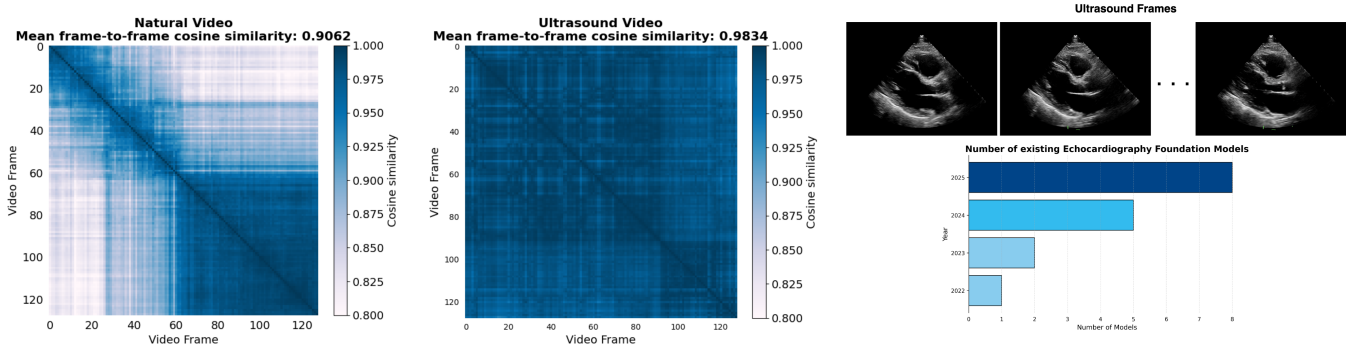


Figure 2: The figure on the left shows frame-level cosine similarity matrices: natural video frames from the SumMe dataset ([Gygli *et al.*, 2014]) versus echocardiography video frames extracted using SigLIP2 ([Tschannen *et al.*, 2025]). Echocardiography videos exhibit much higher frame-to-frame similarity compared to natural videos, making informative feature extraction more challenging. The figure on the right shows the number of echocardiography foundation models released each year.

models, as evidenced by the increasing number of models proposed each year (Figure 2). However, most of these models have been developed and evaluated on private datasets, which makes it difficult to assess their generalizability. ETAB ([M Alaa *et al.*, 2022]) provided an early benchmark with four public datasets, but its tasks overlap across datasets and views, and it focused on supervised adaptation rather than zero-shot or cross-modal evaluation. As a result, no standardized benchmark exists for modern echocardiography foundation models.

Beyond the lack of standardized benchmarks, echocardiography presents unique challenges for foundation model development. Unlike static radiographs, echocardiography produces temporal sequences of 2D images. Ultrasound images are inherently noisy and temporally complex, with high frame-to-frame similarity that makes effective representation learning more difficult ([Kang *et al.*, 2024; Song *et al.*, 2024]). As shown in Figure 2, echocardiography videos exhibit much higher frame-level cosine similarity than natural videos, reflecting the low signal-to-noise ratio and limited visual variation of the modality. These characteristics have been associated with reduced robustness and limited generalization when models are trained directly on noisy ultrasound data ([Javed *et al.*, 2024]). They also make it hard to identify which modeling choices truly help, highlighting the need for systematic evaluation.

Furthermore, it remains unclear how these modality-specific models compare with general-purpose vision foundation models, which have much larger diversity in training data. This raises several fundamental questions: 1) How do echocardiography foundation models perform relative to each other under a fixed evaluation protocol? 2) Are their learned representation spaces fundamentally different from those of general-purpose models, and how do these differences affect downstream tasks? 3) To what extent can they enable zero-shot transfer, and do they exhibit systematic biases across datasets or clinical tasks? Addressing these open questions is essential for establishing reliable foundations for echocardiography AI, with direct implications for both methodological progress and the safe and reliable translation of these tech-

nologies into clinical practice.

This work introduces **CardioBench** (see Figure 1), a comprehensive benchmark for echocardiography foundation models. By unifying eight publicly available datasets into a standardized evaluation suite spanning four regression and five classification tasks, CardioBench establishes the common ground for fair, reproducible, and clinically meaningful comparison. Unlike prior efforts that focused on individual datasets or tasks, CardioBench enables systematic evaluation across functional and structural endpoints, providing a robust basis for tracking progress in this emerging field. It compares leading cardiac-specific models with general-purpose vision and biomedical encoders under consistent zero-shot, probing, and alignment protocols, offering controlled analysis of how architectural design, temporal modeling, and supervision strategies shape transferability. To maximize accessibility and reproducibility, CardioBench provides standardized dataset preprocessing and data splits together with unified evaluation scripts, ensuring that results are directly comparable and easily extendable by the community.

Beyond results, CardioBench provides actionable insights into what drives performance in echocardiography foundation models: the role of temporal modeling, the importance of text encoders, the robustness of retrieval-based methods, and the surprising strengths and weaknesses of generalist backbones. We expect CardioBench to: (1) stimulate the development of new models tailored to the unique challenges of echocardiography, (2) establish a systematic way of measuring model quality for scientific progress, and (3) guide future pretraining strategies by revealing which architectural and supervision choices yield meaningful representations.

2 Related Work

Recent works have advanced benchmarking and foundation models in medical imaging across multiple domains. [Bassi *et al.*, 2024] builds a large-scale segmentation benchmark across nine abdominal organs to test models under distribution shift. Beyond performance, [Jin *et al.*, 2024] emphasized fairness by assessing foundation models across multiple modalities and sensitive attributes. At the same time, [Huix

et al., 2024] highlights the difficulty of transferring general-purpose foundation models to specialized modalities.

In echocardiography, [M Alaa *et al.*, 2022] provided an important early benchmark by assembling four public datasets into 31 tasks, establishing the first standardized protocol for model comparison. Many of these tasks, however, overlap across datasets and views, offering breadth but limited diversity in clinically distinct evaluation endpoints. Furthermore, ETAB primarily focused on supervised task adaptation, leaving a critical gap in evaluating the zero-shot, probing, and cross-modal alignment capabilities of modern multi-modal foundation models. Since then, several echocardiography foundation models have been released, many of which are evaluated only on private datasets, which limits reproducibility and fair comparison across methods ([Song *et al.*, 2024]). Together, these works highlight the absence of a standardized benchmark specifically tailored for echocardiography foundation models, underscoring the need for a public protocol that enables fair evaluation under noise and domain shifts in cardiac ultrasound.

3 Methodology

Cardiobench is a unified benchmark designed to evaluate echocardiography foundation models across multiple clinical tasks, datasets, and evaluation regimes under minimal adaptation. We assess both out-of-the-box performance and lightweight adaptation. We define zero-shot evaluation as inference without any training on the target dataset, using only model components released by the original authors. This setting reflects the intended use of foundation models, which are not retrained for every downstream task or deployment scenario. Together, these design choices provide a consistent framework for comparing models with diverse architectures and training strategies.

Within this framework, we design tasks that reflect the clinical use of echocardiography. Echocardiography provides a comprehensive view of the heart, capturing its motion, structure, and pathological states across time. To rigorously benchmark foundation models in this domain, we design tasks that capture functional, structural, and diagnostic aspects of clinical practice, as illustrated in Figure 1. Functional tasks reflect the heart’s movement over time, with Left Ventricular Ejection Fraction (LV EF) regression serving as a standard measure of global cardiac performance that requires models to capture temporal dynamics across the cardiac cycle. Structural tasks emphasize the heart’s anatomical properties, focusing on diastolic measurements (IVSd, LVId, LVPWd) to assess the spatial localization of cardiac walls. At the same time, diagnostic tasks focus on disease classification, including aortic stenosis (AS), pulmonary arterial hypertension (PAH), atrial septal defect (ASD), ST-elevation myocardial infarction (STEMI), and regional wall motion abnormality (RWMA) from 3 different views, thereby testing adaptability to diverse clinical targets.

Beyond core tasks, the CardioBench also accounts for echocardiography’s broader context, including its multi-view nature and potential demographic biases. Echocardiography is inherently multi-view, with different pathologies and

anatomical structures visible only from specific perspectives. View classification is therefore essential, as accurate recognition enables physicians to interpret the correct structures and ensures that automated models apply the appropriate downstream diagnostic tasks. In addition, we analyze demographic and acquisition-related factors, providing insight into subgroup robustness.

For evaluation, we consider three categories of foundation models: those designed specifically for echocardiography, those trained on broader biomedical data, and large-scale general-purpose models. These span a wide range of architectural choices, from models without text supervision to those with temporal transformers over frame sequences or purely image-level extractors. Taken together, these variations in scale, architecture, and pretraining strategy allow us to assess how different design choices transfer to echocardiography interpretation (see Supplementary Material).

Echocardiography-specific FM. We evaluate the four Echocardiography foundation models with publicly released weights available at the time of writing. The earliest, EchoCLIP ([Christensen *et al.*, 2023]), introduced a vision–language approach to cardiac ultrasound. EchoPrime ([Vukadinovic *et al.*, 2024]) built on this idea with a stronger video encoder and a larger dataset, while also incorporating a separate view classifier and relying on report retrieval at inference time. In parallel, PanEcho ([Holste *et al.*, 2025]) explored an alternative direction by discarding text supervision and instead combining frame features with temporal aggregation in a multitask setup, while EchoFM ([Kim *et al.*, 2024]) explored a generative pretraining strategy centered on reconstructing cardiac motion. As it lacks a text encoder and the weights for its task-specific linear heads were not released, it cannot perform zero-shot evaluation.

Biomedical and general-purpose FM. To assess transfer from broader domains, we also include BioMedCLIP ([Zhang *et al.*, 2023]), a vision–language model pretrained on millions of biomedical image–text pairs spanning radiology, microscopy, pathology, and ultrasound. For comparison, we evaluate two large-scale general-purpose models trained at internet scale: DINOv3 ([Siméoni *et al.*, 2025]), a self-supervised vision transformer, and SigLIP2 ([Tschannen *et al.*, 2025]), a vision–language model designed to produce stronger dense representations. Together, these models enable testing of how far biomedical and generic pretraining can transfer to echocardiography tasks, and whether domain-specific pretraining is required to achieve strong performance.

We design experiments to examine two complementary aspects of foundation models: **(i) the capacity to perform clinically relevant tasks without task-specific training**, and **(ii) the quality of their learned representations for downstream adaptation**. Therefore, we focus on zero-shot evaluation and probing, while excluding fine-tuning and few-shot training, as both are prone to overfitting and require substantial labeled data for stable performance ([Silva-Rodriguez *et al.*, 2024]). Further details on zero-shot evaluation, prompt design, and probing implementations are provided in Supplementary Material.

Foundation models are evaluated on both predictive accuracy and the structure of their learned representations. We

Model	Dynamic	CAMUS	Pediatric	LVH			CardiacNet		HMC-QU	TMED-2	segRWMA		
	LV EF (%) ↓			LVIDd (cm) ↓	IVSd (cm) ↓	LVPWd (cm) ↓	ASD ↑	PAH ↑	STEMI ↑	AS ↑	A2C ↑	A3C ↑	A4C ↑
EchoCLIP [Christensen <i>et al.</i> , 2023]	9.99	<u>9.83</u>	13.80	0.79	0.57	0.41	47.88	46.96	52.51	<u>44.13</u>	35.68	36.27	14.29
EchoPrime [Vukadinovic <i>et al.</i> , 2024] •	<u>7.78</u>	14.00	5.44	-	-	-	-	-	-	<u>44.13</u>	-	-	-
PanEcho [Holste <i>et al.</i> , 2025] •	5.79	11.63	<u>9.10</u>	0.36	0.21	0.18	-	-	-	58.90	30.50	24.30	20.52
BioMedCLIP [Zhang <i>et al.</i> , 2023]	13.83	18.87	18.30	0.97	<u>0.28</u>	0.26	<u>40.24</u>	25.75	33.33	<u>44.13</u>	37.66	32.10	6.67
DINOv3 [Siméoni <i>et al.</i> , 2025]	14.67	9.88	18.24	<u>0.69</u>	<u>0.28</u>	<u>0.22</u>	36.49	<u>41.44</u>	<u>34.21</u>	<u>44.13</u>	47.83	<u>48.00</u>	48.15
SigLIP2 [Tschannen <i>et al.</i> , 2025]	14.66	9.28	18.22	<u>0.69</u>	<u>0.28</u>	<u>0.22</u>	36.49	24.11	32.43	17.38	<u>47.25</u>	72.02	<u>47.17</u>

Table 1: Zero-shot results across 4 regression tasks and 5 classification tasks on 8 publicly available datasets. **Regression tasks** (reported in MAE, lower is better ↓): LV EF on Dynamic [Ouyang *et al.*, 2020], CAMUS [Leclerc *et al.*, 2019], and Pediatric [Reddy *et al.*, 2023]; LVIDd, IVSd, and LVPWd on LVH [Duffy *et al.*, 2022]. **Classification tasks** (reported in F1-macro, higher is better ↑): ASD and PAH on CardiacNet [Yang *et al.*, 2024]; STEMI on HMC-QU [Degerli *et al.*, 2021]; AS on TMED-2 [Huang *et al.*, 2022]; and A2C, A3C, and A4C on segRWMA [Liu *et al.*, 2023]. Models with video-based training are marked with •. The best results are shown in **bold**, and the second best are underlined.

Model	LVH						CardiacNet				HMC-QU		segRWMA					
	LVIDd (cm) ↓	Δ	IVSd (cm) ↓	Δ	LVPWd (cm) ↓	Δ	ASD ↑	Δ	PAH ↑	Δ	STEMI ↑	Δ	A2C ↑	Δ	A3C ↑	Δ	A4C ↑	Δ
EchoCLIP [Christensen <i>et al.</i> , 2023]	0.47	0.32	0.28	0.29	0.22	0.19	38.49	9.39	41.44	5.52	73.99	21.48	47.83	12.15	48.00	11.73	48.15	38.86
EchoPrime [Vukadinovic <i>et al.</i> , 2024] ●	0.41	—	0.25	—	0.19	—	52.66	—	63.36	—	80.00	—	8.33	—	68.48	—	48.15	—
PanEcho [Holste <i>et al.</i> , 2025] ●	0.35	0.01	0.18	0.03	0.15	0.03	58.53	—	61.51	—	69.70	—	72.73	42.23	47.47	23.17	64.78	44.26
EchoFM [Kim <i>et al.</i> , 2024]	0.57	—	0.32	—	0.24	—	50.48	—	41.44	—	71.82	—	47.83	—	48.00	—	48.15	—
BioMedCLIP [Zhang <i>et al.</i> , 2023]	0.52	0.45	0.30	0.02	0.23	0.03	58.53	1.20	41.44	15.69	55.44	22.11	47.83	10.17	48.00	15.90	48.15	41.48
DINOv3 [Siméoni <i>et al.</i> , 2025]	0.47	0.22	0.28	0.00	0.21	0.01	56.76	22.36	58.85	17.41	75.00	40.79	47.83	0.00	48.00	0.00	48.15	0.00
SigLIP2 [Tschannen <i>et al.</i> , 2025]	0.51	0.18	0.30	0.02	0.23	0.01	68.49	32.00	47.96	23.85	75.00	42.57	47.83	0.48	48.00	24.02	48.15	0.98

Table 2: Linear probing results across 3 regression tasks and 4 classification tasks on 4 publicly available datasets. **Regression tasks** (reported in MAE, lower is better ↓): LVIDd, IVSd, and LVPWd on LVH. **Classification tasks** (reported in F1-macro, higher is better ↑): ASD and PAH on CardiacNet; STEMI on HMC-QU; and A2C, A3C, and A4C on segRWMA. Reported Δ values indicate absolute change relative to zero-shot. Models with video-based training are marked with •. The best results are shown in **bold**, and the second best are underlined.

therefore report metrics across four dimensions: task performance, clustering consistency, cross-modal alignment, and demographic robustness. For task performance, we use Mean Absolute Error (MAE) as the primary regression metric and macro-averaged F1 for classification and view classification. Clustering consistency is assessed using the Adjusted Rand Index (ARI), which measures how well embedding clusters recover ground-truth echocardiography views. Cross-modal alignment is evaluated by testing whether visual embeddings align with text prompts. Finally, demographic robustness is examined through subgroup analyses of EF errors stratified by sex, age, BMI, and image quality.

4 Results

We summarize zero-shot performance in Table 1. PanEcho is the most consistent performer, achieving the best or second-best results for ejection fraction (EF) estimation on EchoNet-Dynamic and EchoNet-Pediatric, and outperforming all competitors on the structural regression tasks from EchoNet-LVH. It also leads in classification, with the highest score of 58.90% on TMED-2 aortic stenosis (AS) detection. EchoPrime performs strongly on both regression and classification, which is notable given its retrieval-based inference and the potential influence of similarities between test cases and its private database. A notable observation is the performance of general-purpose foundation models such as SigLIP2 and DINOv3, which deliver strong results despite lacking cardiac-specific pretraining. SigLIP2, in particular, surpasses several specialized echocardiography models on CAMUS EF estimation and achieves competitive performance on segRWMA regional wall abnormality detection. At the same time, both SigLIP2 and DINOv3 perform nearly on

par with PanEcho on EchoNet-LVH regression LVPWd. In classification, they achieve the highest scores in RWMA detection across all three views, even outperforming EchoCLIP, despite EchoCLIP being explicitly trained on cardiac ultrasound. This underperformance is most pronounced on the A4C view, where EchoCLIP lags by more than 34%. Nevertheless, EchoCLIP remains strong on several tasks, achieving F1 scores of 47.88% on ASD and 46.96% on PAH, surpassing the best general-purpose models by margins of 7.61% and 5.52%, respectively. On STEMI detection, it reaches 52.51%, representing an improvement of 18.3% over competitors.

The linear probing performance is summarized in Table 2. On regression tasks, PanEcho maintains a clear advantage, achieving the lowest errors across all EchoNet-LVH measurements (MAE of 0.35 on LVIDd, 0.15 on IVSd, and 0.30 on LVPWd), with only marginal improvements from linear probing ($\Delta \leq 0.03$). By contrast, general-purpose encoders such as DINOv3 and SigLIP2 show larger reductions in error (0.20–0.23 MAE), narrowing the gap to PanEcho, though they remain behind. These results illustrate that EchoNet-LVH structural regression benefits less from probing. For classification, linear probing yields more pronounced changes. SigLIP2 improves by 32% on ASD to reach 68.49% F1, outperforming all specialized models by nearly 10%. On PAH and STEMI, however, EchoPrime delivers the strongest performance, achieving 63.36% and 80.00%, while SigLIP2 remains competitive at 47.96% and 72.57%, respectively. These results show that general-purpose encoders can not only close the gap but, in some cases, even surpass specialized models.

In RWMA detection, PanEcho achieves the highest gains, with improvements of 42.23% on A2C and 44.26% on A4C, reaching 72.73% and 64.78%, respectively. EchoPrime ex-

Model	LVH $\uparrow$	CardiacNet $\uparrow$	CAMUS $\uparrow$ ●	Dynamic $\uparrow$ ●	Pediatric $\uparrow$ ●	HMC-QU $\uparrow$ ●	TMED-2 $\uparrow$ ●	segRWMA $\uparrow$ ●
EchoCLIP [Christensen <i>et al.</i> , 2023]	1.76	27.12	33.11	8.55	20.95	34.33	14.25	16.86
EchoPrime [Vukadinovic <i>et al.</i> , 2024]	98.66	<u>34.59</u>	16.39	98.49	79.53	88.19	62.86	15.79
BioMedCLIP [Zhang <i>et al.</i> , 2023]	0.57	76.11	17.02	26.37	18.41	<u>47.67</u>	<u>21.98</u>	18.41
DINOV3 [Siméoni <i>et al.</i> , 2025]	0.00	0.00	0.00	0.31	35.82	0.00	4.89	0.00
SigLIP2 [Tschannen <i>et al.</i> , 2025]	<u>29.05</u>	8.75	57.01	<u>87.29</u>	<u>45.32</u>	41.37	16.17	2.43

Table 3: View classification results across 8 publicly available datasets, reported in **F1-macro** score. Multi-view datasets are marked with ●. The best results are shown in **bold**, and the second best are underlined.

cels on A3C, where it reaches 68.48%, while EchoCLIP remains flat at 48.00% across all views, converging with DINOV3 and SigLIP2 despite its cardiac-specific training. Overall, linear probing highlights complementary strengths with PanEcho remaining unrivaled on regression and two RWMA views, EchoPrime achieving the best results on PAH and STEMI, and SigLIP2 surpassing all competitors on ASD.

View classification results in Table 3 show that EchoPrime achieves the highest F1 scores on the majority of datasets, benefiting from its supervised, pretrained view classifier rather than relying solely on text-prompt alignment. It leads on five out of eight datasets, highlighting the strength of its dedicated view recognition module. Interestingly, the remaining datasets are topped by models without cardiac-specific pretraining: BioMedCLIP achieves the best results on CardiacNet (76.11%) and TMED-2 (62.86%), while SigLIP2 outperforms all others on CAMUS (57.01%). For general-purpose models, however, view classification can break down when learned representations do not sufficiently separate visually similar echocardiographic views, leading to near-zero macro-F1 on some datasets. EchoCLIP, despite being trained specifically on echocardiography, fails to dominate on any dataset and often lags behind BioMedCLIP or general-purpose models. These findings suggest that while supervised view classifiers provide a clear advantage, large-scale pretraining on diverse medical or natural images can transfer surprisingly well to echocardiography view classification.

Subgroup analyses reveal distinct biases in EF estimation on CAMUS that are less pronounced in EchoNet-Pediatric (Figure 3), despite overall performance trends being consistent across models. On CAMUS (Figure 3a,b,c), subgroup differences are evident: younger patients (≤ 45) and scans labeled as “Good” quality show larger errors and wider spreads, likely reflecting distribution biases since most samples fall into the “Medium” quality category, where models perform best. A modest sex gap is also visible, with females showing slightly higher errors, particularly for EchoPrime and PanEcho. In the larger EchoNet-Pediatric cohort (Figure 3d,e,f), these disparities are less pronounced. Sex- and age-related differences largely disappear, while BMI exhibits the expected trend: healthy ranges yield lower errors, whereas both low and high extremes increase variability, consistent with the physics of ultrasound imaging, where excessive or insufficient tissue layers can degrade acoustic penetration and image quality. Across both datasets, SigLIP2 and DINOV3 maintain the most stable performance across demographic and acquisition subgroups, showing narrow error distributions and minimal subgroup-related shifts. BioMedCLIP, while consistently higher in absolute error, also shows relatively uni-

form behavior across subgroups. By contrast, PanEcho and EchoPrime demonstrate more outliers and wider error distributions across several subgroups, particularly in females and younger patients on CAMUS and in BMI extremes on EchoNet-Pediatric.

5 Discussion

CardioBench reveals that no single foundation model dominates across all tasks, datasets, and evaluation regimes. Instead, performance depends strongly on the interaction between model design choices, dataset characteristics, and evaluation setup.

Modeling EF regression. PanEcho and EchoPrime stand apart from the contrastive approaches in CardioBench because their zero-shot predictions are not driven by text encoders. PanEcho leverages its multitask design to achieve the lowest errors on EchoNet-Dynamic and strong results on Pediatric, showing that supervised EF knowledge can transfer effectively across datasets. EchoPrime, in contrast, benefits from retrieval: rather than modeling EF as a smooth continuum, it assigns labels by matching test cases to similar exemplars in its joint space. This discrete matching helps on EchoNet-Pediatric, where it outperforms contrastive models, but the approach fails on CAMUS, where scanner heterogeneity may distort embeddings and make nearest-neighbor matches unreliable. Both models incorporate temporal dynamics, but differ in how strongly their predictions depend on them. A frame-shuffling stress test highlights this contrast: shuffling frames increases PanEcho MAE from 5.79 to 6.32, indicating sensitivity to temporal order, whereas EchoPrime remains relatively stable (7.79 to 6.12), suggesting that its retrieval mechanism can fall back on exemplar similarity even when sequence order is disrupted.

To examine contrastive approaches, we directly assess whether they encode EF as a cross-modal dimension. We construct a text axis from prompts spanning 0–100% EF, normalize these embeddings, and extract the first principal component. Visual embeddings from test videos are then projected onto this axis, and their Pearson correlation with ground-truth EF quantifies alignment (Figure 12, Supplementary Material). This analysis reveals significant differences between models. EchoCLIP, trained on cardiac ultrasound reports, is the only model to recover a physiologically meaningful EF axis ($r = 0.52$ on EchoNet-Dynamic, $r \approx 0.2$ – 0.3 on CAMUS and Pediatric), suggesting that domain-specific text encoders can enforce monotonic cross-modal structure. BioMedCLIP, despite pretraining on extensive biomedical corpora, shows almost no alignment ($r \approx 0$), indicating that general medical semantics are insufficient to ground EF

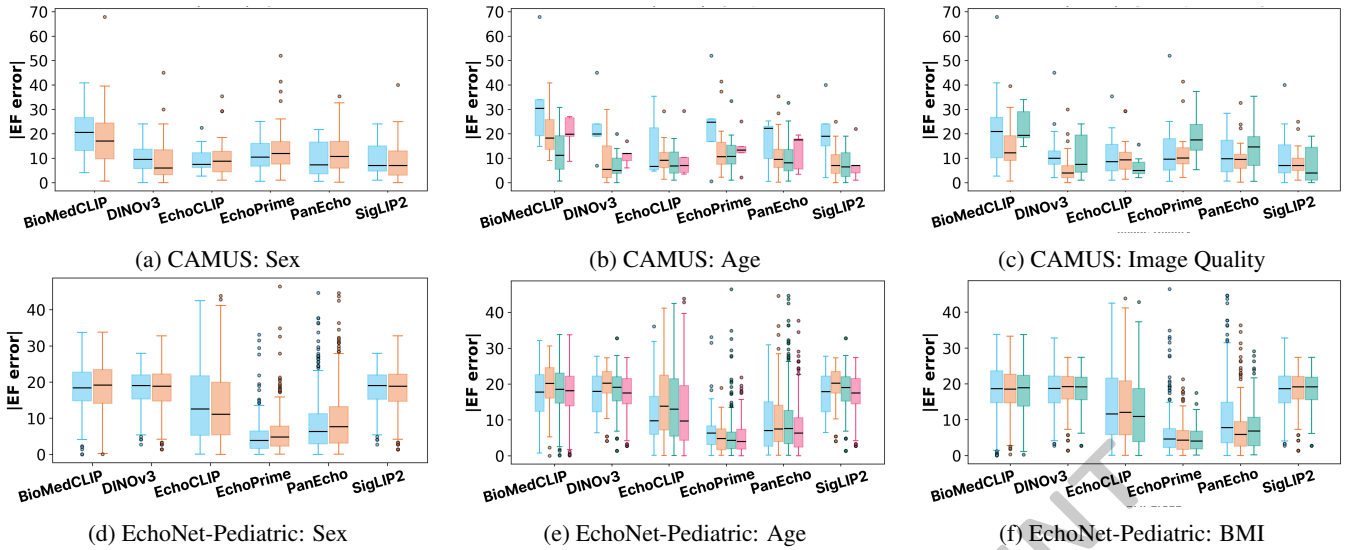


Figure 3: Distribution of absolute EF prediction error across demographic and image-quality subgroups. Top row: CAMUS stratified by sex, age (blue: < 45 , orange: $46\text{--}65$, green: $66\text{--}80$, pink: > 80), and image quality (blue: good, orange: medium, green: poor). Bottom row: EchoNet-Pediatric stratified by sex (blue: female, orange: male), age (blue: $0\text{--}1$, orange: $1\text{--}5$, green: $6\text{--}12$, pink: $13\text{--}18$), and BMI (blue: underweight, orange: healthy weight, green: overweight/obese).

as a continuous variable. General-purpose models such as SigLIP2 and DINOv3 also result in near-zero correlations, yet achieve their strongest results on CAMUS. At first glance, this might suggest robustness to acquisition shifts; however, a closer look indicates that these gains are not physiologically grounded. Specifically, we observe that SigLIP2 achieves lower MAE on images with poor quality compared to those of higher quality (Figure 3c), which is counterintuitive from a clinical perspective. This pattern suggests that the apparent success of generalist models on CAMUS reflects sensitivity to dataset-specific artifacts rather than meaningful encoding of EF, explaining their poor generalization outside this narrow setting.

Clustering challenges in view classification. A similar picture emerges in view classification, where architectural choices again dominate over text alignment. EchoPrime achieves the strongest results across multiple datasets by leveraging its supervised view head, demonstrating that explicitly modeling clinical structure can result in zero-shot advantages. By contrast, EchoCLIP struggles to generalize beyond A4C despite being trained on this view, because its contrastive objective emphasizes alignment with reports rather than enforcing consistent view identity. As a result, its embeddings entangle clinical content with anatomical cues, limiting transfer even on its main training view. Large-scale encoders such as BioMedCLIP and SigLIP2 occasionally outperform specialized models on datasets like EchoNet-Pediatric and CAMUS (Figure 14, Supplementary Material), but UMAP projections (Figure 4) of TMED-2 embeddings reveal that none of the models form globally distinct view clusters. Interestingly, BioMedCLIP, EchoCLIP, and PanEcho, which were not explicitly trained for view classification, tend to group PLAX and PSAX together while mixing A2C and A4C, as these views are indeed visually similar

within short-axis and long-axis families. kNN probing (Table 4) recovers some discriminative power, ranking BioMedCLIP highest, followed by PanEcho and EchoCLIP, while SigLIP2 surpasses EchoPrime when its supervised view classifier is removed. This shows that EchoPrime’s advantage comes almost entirely from its explicit classifier head, while other models contain partial view information in their embeddings that kNN can recover locally, but which does not form globally distinct clusters or generalize consistently across datasets.

Embedding structures for pathology tasks. Within CardioBench, inspection of the embedding spaces for classification tasks evidences that zero-shot performance is constrained by weakly discriminative representation spaces. As shown in the UMAP visualizations in Figure 13 (Supplementary Material), pathology-present and pathology-absent cases form partially separable but substantially overlapping clusters, with limited intra-class compactness and low silhouette scores across datasets. This indicates the limited prioritization of pathology-specific cues in current visual backbones, which tend instead to encode broader distributional features. The contrast with linear probing, showing substantially higher performance for BioMedCLIP and SigLIP2, further highlights that discriminative signals are present but not aligned with text prompts or directly accessible for zero-shot. These findings underscore the gap between latent signal and usable representation, emphasizing the need for models that organize clinical information more explicitly.

CardioBench makes clear that progress in echocardiography foundation models cannot be measured by zero-shot performance alone. Across regression, classification, and view recognition, the benchmark reveals a consistent pattern: models contain latent clinical signal, but its accessibility depends heavily on architectural design, training supervision, and the

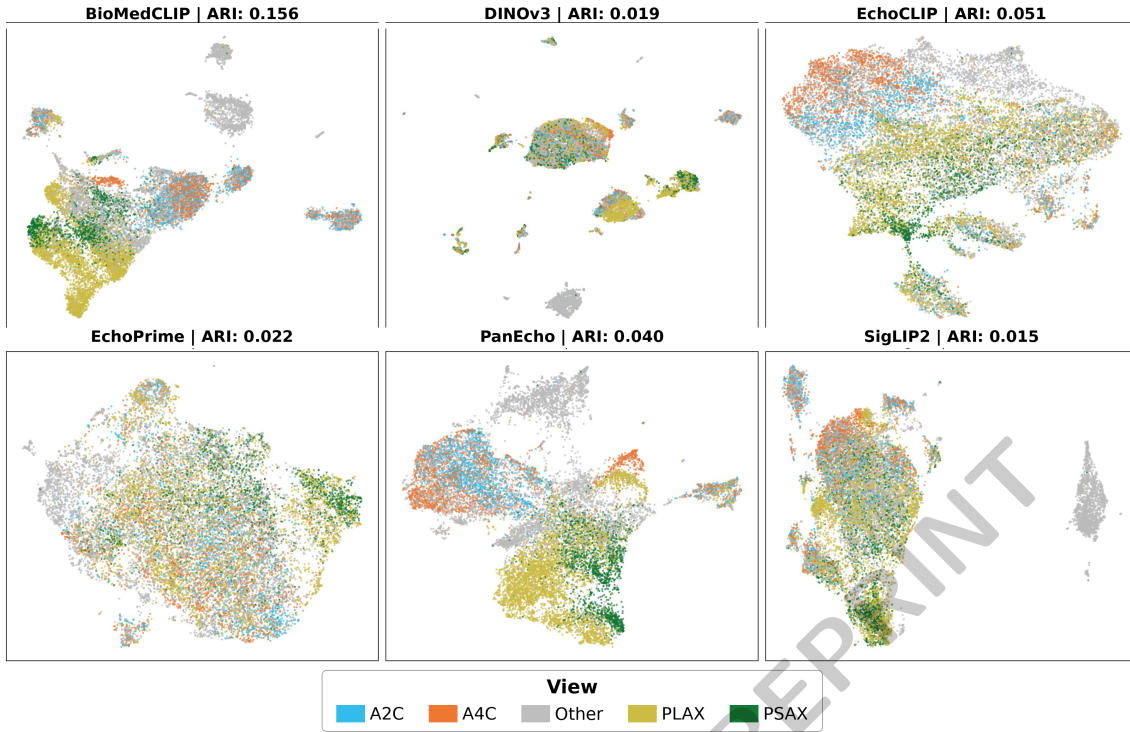


Figure 4: UMAP projection of TMED-2 embeddings highlighting view-based clustering.

Model	F1	Δ	A2C	Δ	A4C	Δ	PLAX	Δ	PSAX	Δ	Oth.	Δ
EchoCLIP	59.98	45.7	53.9	0.0	48.1	19.1	62.7	27.5	84.6	54.1	49.1	45.6
EchoPrime	50.87	12.0	37.7	49.2	39.8	46.0	70.1	15.4	66.6	32.6	33.2	10.1
PanEcho	63.46	–	57.6	–	47.7	–	66.0	–	84.1	–	70.2	–
BioMedCLIP	68.13	41.8	59.6	54.6	57.2	5.8	71.4	42.5	88.6	74.5	66.3	60.6
DINOv3	50.92	46.0	36.7	29.3	43.3	42.8	61.0	61.0	75.9	25.5	40.0	39.8
SigLIP2	57.34	41.2	42.1	1.7	47.4	30.0	69.7	67.3	79.1	78.6	42.6	25.6

Table 4: KNN probing results on TMED-2, reporting F1-macro scores and per-view accuracies. Δ denotes the absolute change relative to zero-shot performance. The best results are **bold**, and the second-best are underlined.

stability of the embedding organization. This points to several practical directions. First, explicit supervision for core clinical axes such as EF or view classification proves more reliable than expecting them to emerge implicitly, suggesting that pretraining pipelines should integrate lightweight but structured supervision. Second, temporal modeling is indispensable for functional tasks, as demonstrated by PanEcho, while retrieval-based matching offers complementary robustness, motivating hybrid approaches that combine the strengths of both. Third, domain-specific text encoders, as in EchoCLIP, can enforce physiologically meaningful cross-modal structure, but their advantage is not stable, underscoring the need to broaden cardiac text corpora. Finally, the surprisingly strong performance of general-purpose encoders such as SigLIP2 and DINOv3 highlights both an opportunity and a limitation: scale and diversity alone can produce robust baselines under domain shift, yet these models fail to organize clinical signals in a way that supports fine-grained reasoning. This suggests that future cardiac foundation models

should not discard generalist architectures, but rather adapt them through targeted supervision and domain grounding, bridging the gap between robustness and clinical fidelity.

6 Conclusion

CardioBench demonstrates that echocardiography foundation models must be assessed through multi-evaluation to capture their true capabilities. Performance depends on design and supervision choices shaping temporal dynamics, retrieval, and clinically grounded representations. Future advances will likely come from hybrid approaches combining these strengths. By providing a publicly available standardized benchmark, CardioBench establishes a baseline for fair comparison to develop clinically meaningful models.

Contribution Statement

Darya Taratynova and Ahmed Aly contributed equally to this work. Numan Saeed and Mohammad Yaqub provided joint supervision.

References

- [Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- [Bassi *et al.*, 2024] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian R Rokuss, Ziyang Huang, et al. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Advances in Neural Information Processing Systems*, 37:15184–15201, 2024.
- [Christensen *et al.*, 2023] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Multimodal foundation models for echocardiogram interpretation. *arXiv preprint arXiv:2308.15670*, 2023.
- [Degerli *et al.*, 2021] Aysen Degerli, Morteza Zabihi, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, Ridha Hamila, and Moncef Gabbouj. Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access*, 9:34442–34453, 2021.
- [Duffy *et al.*, 2022] 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, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning, 2022.
- [Ghiasi *et al.*, 2022] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pages 540–557. Springer, 2022.
- [Gygli *et al.*, 2014] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014.
- [Holste *et al.*, 2025] Gregory Holste, Evangelos K Oikonomou, Márton Tokodi, Attila Kovács, Zhangyang Wang, and Rohan Khera. Panecho: Complete ai-enabled echocardiography interpretation with multi-task deep learning. *medRxiv*, pages 2024–11, 2025.
- [Huang *et al.*, 2022] Zhe Huang, Gary Long, Benjamin Wessler, and Michael C Hughes. Tmed 2: a dataset for semi-supervised classification of echocardiograms. In *In DataPerf: Benchmarking Data for Data-Centric AI Workshop*, 2022.
- [Huang *et al.*, 2023] Ziyang Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023.
- [Huix *et al.*, 2024] Joana Palés Huix, Adithya Raju Ganesan, Johan Fredin Haslum, Magnus Söderberg, Christos Matsoukas, and Kevin Smith. Are natural domain foundation models useful for medical image classification? In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 7634–7643, 2024.
- [Irvin *et al.*, 2019] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpan-skaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.
- [Javed *et al.*, 2024] Haseeb Javed, Shaker El-Sappagh, and Tamer Abuhrmed. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust ai applications. *Artificial Intelligence Review*, 58(1):12, 2024.
- [Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [Jin *et al.*, 2024] Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, DOU QI, S Kevin Zhou, and Xiaoxiao Li. Fairmedfm: fairness benchmarking for medical imaging foundation models. *Advances in Neural Information Processing Systems*, 37:111318–111357, 2024.
- [Johnson *et al.*, 2019] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- [Kang *et al.*, 2024] Hyekyoung Kang, Chanrok Park, and Hyungjin Yang. Evaluation of denoising performance of resnet deep learning model for ultrasound images corresponding to two frequency parameters. *Bioengineering*, 11(7):723, 2024.
- [Kim *et al.*, 2024] Sekeun Kim, Pengfei Jin, Sifan Song, Cheng Chen, Yiwei Li, Hui Ren, Xiang Li, Tianming Liu, and Quanzheng Li. Echofm: Foundation model for generalizable echocardiogram analysis. *arXiv preprint arXiv:2410.23413*, 2024.
- [Leclerc *et al.*, 2019] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.
- [Li *et al.*, 2022] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

- [Liu *et al.*, 2023] Kaikai Liu, Yiyu Shi, Jian Zhuang, Meiping Huang, Hongwen Fei, Boyang Li, Jin Hong, Qing Lu, Erlei Zhang, and Xiaowei Xu. Enhance regional wall segmentation by style transfer for regional wall motion assessment. In *BMVC*, pages 462–466, 2023.
- [M Alaa *et al.*, 2022] Ahmed M Alaa, Anthony Philippakis, and David Sontag. Etab: A benchmark suite for visual representation learning in echocardiography. *Advances in Neural Information Processing Systems*, 35:19075–19086, 2022.
- [Mitchell *et al.*, 2019] Carol Mitchell, Peter S Rahko, Lori A Blauwet, Barry Canaday, Joshua A Finstuen, Michael C Foster, Kenneth Horton, Kofo O Ogunyankin, Richard A Palma, and Eric J Velazquez. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the american society of echocardiography. *Journal of the American Society of Echocardiography*, 32(1):1–64, 2019.
- [Ouyang *et al.*, 2020] 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, et al. Video-based ai for beat-to-beat assessment of cardiac function, 2020.
- [Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [Reddy *et al.*, 2023] Charitha D Reddy, Leo Lopez, David Ouyang, James Y Zou, and Bryan He. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients, 2023.
- [Roy *et al.*, 2023] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of conynets for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–415. Springer, 2023.
- [Silva-Rodriguez *et al.*, 2024] Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23681–23690, 2024.
- [Siméoni *et al.*, 2025] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.
- [Singh *et al.*, 2022] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022.
- [Song *et al.*, 2024] Ke Song, Jing Feng, and Duo Chen. A survey on deep learning in medical ultrasound imaging. *Frontiers in Physics*, 12:1398393, 2024.
- [Tschannen *et al.*, 2025] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.
- [Vukadinovic *et al.*, 2024] Milos Vukadinovic, Xiu Tang, Neal Yuan, Paul Cheng, Debiao Li, Susan Cheng, Bryan He, and David Ouyang. Echoprime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation. *arXiv preprint arXiv:2410.09704*, 2024.
- [Yang *et al.*, 2024] Jiewen Yang, Yiqun Lin, Bin Pu, Jiarong Guo, Xiaowei Xu, and Xiaomeng Li. Cardiacnet: Learning to reconstruct abnormalities for cardiac disease assessment from echocardiogram videos. In *European Conference on Computer Vision*, pages 293–311. Springer, 2024.
- [Zhang *et al.*, 2023] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.