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

Assessing Spatio-temporal Foundation Models for Clinical Mitral Regurgitation Grading

Jonas Fallmann[1], Martin Schobesberger[2], Volodymyr Tretiak[2], Ilinca Damian[6], Cornelia Roschger[6], Jörg Kellermair[7,8], Clemens Steinwender[7,8], Andreas Zierer[6,8], Alois Ferscha[2], and Erich Kobler[1,3,4,5]

[1] Institute for Machine Learning, Johannes Kepler University, Linz, Austria
[2] Institute of Pervasive Computing, Johannes Kepler University, Linz, Austria
[3] LIT AI Lab, Johannes Kepler University, Linz, Austria
[4] Department of Virtual Morphology, Johannes Kepler University, Linz, Austria
[5] Clinical Research Institute Medical AI, Johannes Kepler University, Linz, Austria
[6] Department of Cardiac-, Vascular- and Thoracic Surgery, Kepler University Hospital, Johannes Kepler University, Linz, Austria
[7] Department of Cardiology and Medical Intensive Care, Kepler University Hospital, Johannes Kepler University, Linz, Austria
[8] Clinical Research Institute for Cardiovascular and Metabolic Diseases, Johannes Kepler University, Linz, Austria

Abstract. Mitral regurgitation (MR) is a common valvular heart disease requiring accurate echocardiographic assessment, yet grading from color Doppler echocardiography remains challenging due to patient variability, imaging artifacts, and the need to interpret complex spatio-temporal flow patterns. In this work, we systematically evaluate recent spatio-temporal foundation models for MR severity grading using apical four-chamber echocardiographic videos. To this end, we compare contrastive, reconstructive, and latent-predictive pre-training paradigms across echocardiography-specific and natural video pre-training domains, including EchoCLIP, EchoFM, EchoJEPA, and V-JEPA variants, alongside a fully supervised 3D CNN baseline. We consider both a four-class grading task and a binary clinically significant MR triage task. Models are evaluated using frozen encoders with linear and attentive probing to assess global and localized representation quality. Across a retrospectively curated dataset of 413 videos, attentive probing consistently improves performance over linear evaluation, highlighting the importance of localized spatio-temporal aggregation. Latent-predictive JEPA-based models achieve the strongest performance among foundation models, approaching or surpassing the fully supervised baseline, while natural video pre-training remains competitive with domain-specific approaches.

Keywords: Mitral Regurgitation · Spatio-temporal Foundation Models · Image Classification.

1 Introduction

Mitral regurgitation (MR) is a disorder of the mitral valve, the anatomical structure that regulates blood flow between the left atrium and left ventricle. During normal cardiac function, the mitral valve closes during systole to prevent retrograde blood flow into the left atrium. Structural or functional abnormalities of the valve apparatus can impair leaflet coaptation, resulting in regurgitation of blood from the left ventricle back into the left atrium [4]. MR is among the most common valvular heart diseases worldwide, affecting millions of patients and representing a growing healthcare burden in ageing populations [5]. Progressive MR induces chronic volume overload, adverse cardiac remodelling, elevated filling pressures, and is associated with increased risks of heart failure, atrial fibrillation, and sudden cardiac death [15].

Echocardiography (EC) is the clinical gold standard for diagnosing and monitoring MR because it provides real-time visualization of both cardiac anatomy and hemodynamics. In particular, color Doppler imaging enables the assessment of blood flow velocity and direction, allowing regurgitant jets to be visualized and graded according to established clinical guidelines [19, 6]. The apical four-chamber (A4C) view plays a central role in MR assessment, as the regurgitant jet is typically aligned with the ultrasound beam, facilitating accurate velocity measurements and semi-quantitative severity grading. Despite its widespread use, EC interpretation remains a challenging and highly operator-dependent task due to inter-patient variability, imaging artifacts, and frequently limited image quality.

Recent advances in artificial intelligence have shown considerable promise for automated EC analysis, including view classification, cardiac function assessment, and valvular disease detection. Existing approaches for MR grading primarily rely on supervised learning methods trained on small-scale task-specific datasets [1, 8, 18, 9, 16, 10]. While these approaches can achieve strong predictive performance, they require substantial annotation effort and often learn representations tailored to a single downstream task. In contrast, foundation models trained through self-supervised learning on large-scale video datasets have recently emerged as a powerful alternative for learning transferable visual representations. Several EC-specific foundation models have been proposed [3, 7, 13], alongside increasingly capable general-purpose video foundation models [2, 14], yet their utility for valvular heart disease assessment remains largely unexplored.

In this work, we systematically evaluate self-supervised video foundation models for MR assessment from color Doppler A4C EC videos. We compare contrastive, reconstructive, and latent-predictive pre-training paradigms across EC-specific and general-purpose video models using linear and attentive probing. We further disentangle the effects of pre-training *objective* and *domain* by contrasting EC-specific with natural video representations. On a clinically curated cohort of 361 patients, we provide the first comparative study of foundation model representations for MR grading and show that latent-predictive JEPA-based models yield the most transferable representations for both fine-grained MR grading and clinically significant MR classification.

Table 1. Foundation models evaluated in this study.

Model	Params	Pre-training Domain	Pre-training Objective
EchoCLIP	151M	EC	Contrastive (vision–language)
EchoFM	300M	EC	Reconstructive (MAE)
EchoJEPA-L	300M	EC	Latent-predictive (JEPA)
EchoJEPA-G	1B	EC	Latent-predictive (JEPA)
V-JEPA 2	1B	Natural Video	Latent-predictive (JEPA)
V-JEPA 2.1	1B	Natural Video	Latent-predictive (JEPA)

2 Related work

Deep Learning for MR Grading. Automated MR assessment from echocardiography (EC) spans both video-based and structured-data approaches. For structured data, classical machine learning applied to EC parameters has yielded binary classification accuracies up to 0.97 [16]. More commonly, video-based deep learning—such as R2+D1 CNNs [18] and color Doppler classifiers [9]—has demonstrated strong diagnostic performance across various institutional cohorts (such as Stanford, UCSF, and Cornell). Recent video-based models frequently achieve F1 scores exceeding 0.90 [8, 9], multi-class accuracies above 80% [10], and AUCs up to 0.97 [1]. However, while these supervised methods perform strongly on their target distributions, they remain fundamentally limited by their reliance on large volumes of task-specific annotations and their poor cross-dataset transferability.

Video and EC Foundation Models. Recent foundation models for echocardiographic video can be organized along two axes: the *pre-training objective* and the *pre-training domain* (see Table 1). This decomposition motivates our central question of which self-supervised objective yields the most transferable representations and whether domain-specific pre-training is required.

EC-specific models: EchoCLIP [3] is a vision-language model trained via contrastive alignment between echocardiogram videos and associated reports; only its vision encoder is used in the subsequent analysis. EchoFM [7] follows a masked autoencoding paradigm, reconstructing pixel-level content from masked patches and thereby also modeling ultrasound artefacts such as speckle. In contrast, EchoJEPA [13] adopts a latent-predictive objective that predicts representations of masked regions, encouraging invariance to low-level artefacts while preserving clinically relevant structure such as chamber geometry and motion.

JEPA and pre-training domain: EchoJEPA [13] shares its architecture and learning principle with V-JEPA 2 [2] and V-JEPA 2.1 [14], which are pre-trained on natural video data; V-JEPA 2.1 further introduces dense prediction over both visible and masked tokens. Because EchoJEPA and V-JEPA models differ only in pre-training domain, they provide a controlled setting to isolate the effect of in-domain versus out-of-domain pre-training.

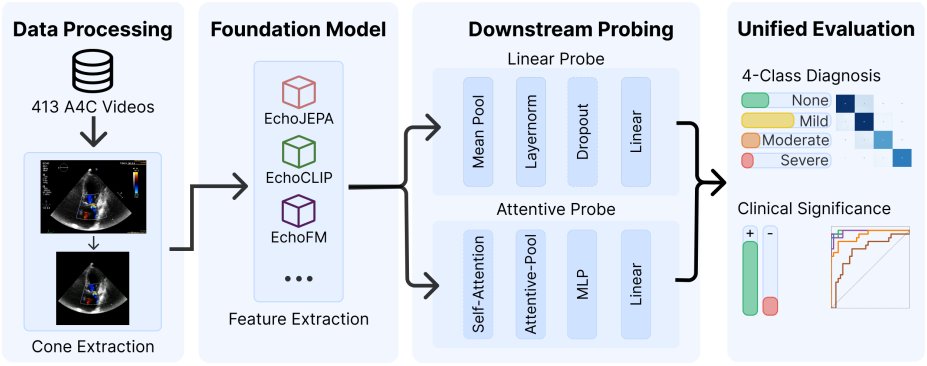

Fig. 1. Foundation model evaluation pipeline. A dataset of 413 A4C videos undergoes data processing to isolate the echo cone. Next, frozen spatio-temporal foundation models extract latent features, which are fed into linear or attentive probes. Unified evaluation is performed across two MR classification tasks: grading (None, Mild, Moderate, Severe) and clinical significance estimation.

3 Methodology

3.1 Clinical Dataset and Cohort

The dataset used in this study was retrospectively curated by a cardiac surgeon and a cardiologist from routine mitral regurgitation (MR) assessments performed between 2016 and 2021. Cases with diagnostically sufficient A4C color Doppler views were selected from a clinical archive of >90,000 echocardiograms. Each patient was previously graded according to established MR severity guidelines into one of four classes (MR 0–III; none to most severe) [17].

The final anonymized cohort comprised 361 patients (222 male, 135 female, 4 no-gender provided) with a mean age of 69.3 ± 13.6 years (range 20–98). Class distributions were 108, 112, 68, and 73 patients for MR grades 0, I, II, and III, respectively.

3.2 Evaluation Framework

To evaluate the clinical utility of foundation models for mitral regurgitation diagnosis, we employ a comprehensive probing pipeline (Figure 1). We benchmark six frozen video models – spanning contrastive, reconstructive, and latent-predictive objectives across both echocardiography and general domains (Table 1) – against a fully supervised baseline. To ensure a fair evaluation of architectural differences independent of model capacity, we specifically trained EchoJEPA-L, a smaller 300-million parameter variant. This allows for a direct, equivalently sized comparison against EchoFM and EchoCLIP while including the larger JEPA variants to evaluate the effects of both scaling and domain vs. general purpose pre-training.

We assess representational quality across two downstream tasks: (1) a fine-grained four-class grading task (None [0], Mild [I], Moderate [II], Severe [III]), and (2) a binary clinical triage task separating *clinically non-significant* (0, I) from *clinically significant* (II, III) cases. Evaluating both tasks provides a holistic measure of the models' expressivity, decoupling nuanced differentiation from actionable clinical utility.

3.3 Probing Architectures and Baseline

As diverse model architectures provide different embedding shapes, we standardize all model outputs into a unified sequence format $\mathrm{N} \times \mathrm{D}$ where N is the number of tokens and D the embedding dimension. This design accommodates varying output structures: patch-level spatio-temporal embeddings from our EchoJEPA variants, V-JEPA, and EchoFM, as well as frame-level embeddings from EchoCLIP. For evaluation, we freeze the respective encoder weights and train linear and attentive probes on top of the last layer latent features. Although prior work suggests MAEs like EchoFM benefit from intermediate representations, probing the latter half of its layers yielded only marginal gains in our setup (see Supplementary Table S1). Thus, we maintained a uniform evaluation protocol across all architectures by extracting only the final layer.

For Linear probing, we perform mean pooling along the sequence dimension (N), collapsing video representations to $\mathbb{R}^{\mathrm{D}}$. This vector is then passed through Layer Normalization, optional dropout, and a single linear classification layer.

Because MR is highly localized in both space and time, we also train an attentive probe to selectively weight clinically relevant features. The token sequence is optionally passed through a stack of self-attention blocks before applying an attentive pooling operation. Using a single learned query for cross-attention, the most important spatio-temporal representation $\mathbb{R}^{\mathrm{D}}$ is dynamically extracted and subsequently processed by a linear classification head.

To establish a baseline against traditional fully supervised methods, we utilize a ResNet-(2+1)D model initialized with pre-trained Kinetics-400 weights, which is then fine-tuned using the raw video sequences. By explicitly factorizing standard 3D convolutions into distinct 2D spatial and 1D temporal operations, this architecture effectively reduces computational complexity while maintaining robust spatio-temporal modeling capabilities.

4 Experiments and Results

4.1 Experimental Setup

Models are evaluated on the 413 video samples using 5×4 nested cross-validation, strictly grouped by patient to prevent data leakage and stratified by diagnosis. Foundation model inputs comprise 128 frames (224×224), split into 32-frame segments. Following pre-training configurations, we use a 16×16 patch size and a tubelet size of 2 for JEPA variants and 4 for EchoFM. Embedded segments are concatenated prior to probing.

Table 2. Overview of performance metrics in the 4-class diagnosis setting. Parameters represent probe without backbone. Best results per metric are written in **bold**, second best underlined.

	Model	Params	F1	ROC-AUC	Specificity	Sensitivity	Precision	NPV
	3D CNN	30M	.642 ± .071	**.875** ± .057	.870 ± .026	.648 ± .070	.645 ± .071	.877 ± .028
Attentive	EchoCLIP	2.9M	.629 ± .045	.837 ± .030	.867 ± .013	.632 ± .041	.648 ± .044	.867 ± .015
	EchoFM	11.5M	.341 ± .039	.628 ± .039	.775 ± .017	.387 ± .044	.319 ± .038	.795 ± .020
	EchoJEPA-L	11.5M	**.665** ± .059	.873 ± .036	**.876** ± .020	**.668** ± .060	**.684** ± .055	**.882** ± .022
	EchoJEPA-G	21.8M	.654 ± .056	.865 ± .032	.870 ± .012	.654 ± .056	.671 ± .068	.876 ± .017
	V-JEPA 2	21.8M	.630 ± .069	.866 ± .049	.863 ± .020	.637 ± .069	.644 ± .076	.871 ± .027
	V-JEPA 2.1	21.8M	.476 ± .066	.780 ± .037	.805 ± .026	.512 ± .062	.545 ± .076	.836 ± .021
Linear	EchoCLIP	3076	.551 ± .069	.787 ± .032	.838 ± .030	.554 ± .067	.560 ± .070	.838 ± .024
	EchoFM	6148	.333 ± .050	.590 ± .031	.771 ± .020	.382 ± .059	.324 ± .054	.794 ± .027
	EchoJEPA-L	6148	.507 ± .056	.809 ± .024	.834 ± .025	.529 ± .034	.545 ± .042	.838 ± .012
	EchoJEPA-G	8452	.517 ± .047	.791 ± .038	.822 ± .013	.537 ± .036	.552 ± .084	.841 ± .037
	V-JEPA 2	8452	.529 ± .087	.791 ± .052	.835 ± .029	.548 ± .091	.542 ± .102	.839 ± .010
	V-JEPA 2.1	8452	.331 ± .054	.717 ± .026	.746 ± .021	.402 ± .040	.412 ± .098	.808 ± .025

All training procedures – including both the probes and the 3D CNN baseline – minimize the cross-entropy loss using `bfloat16` mixed precision, the AdamW optimizer [11],and a cosine learning rate scheduler without warmup. Hyperparameters are optimized within the inner loop via grid search over learning rate η and weight decay γ. The attentive probes, using no additional self-attention blocks, and the 3D CNN baseline share the same search space: $\eta \in \{10^{-3}, 5 \cdot 10^{-4}, 10^{-4}\}$ and $\gamma \in \{4 \cdot 10^{-1}, 10^{-1}, 10^{-2}, 10^{-4}\}$. Linear probes, trained without dropout, are optimized over $\eta \in \{10^{-2}, 5 \cdot 10^{-3}, 10^{-3}\}$ and $\gamma \in \{0, 10^{-2}, 10^{-4}\}$. Attentive probes, linear probes, and the 3D CNN baseline are trained for 15, 12, and 10 epochs, respectively. While foundation models use their native pipelines, the 3D CNN baseline specifically utilizes Kinetics-400 normalization alongside spatial augmentations, namely resizing to 255×255 followed by random 224×224 cropping and horizontal flipping.

Final models are retrained on the outer folds using the optimal inner-loop hyperparameters. We report support-weighted F1-score, ROC-AUC, precision, sensitivity, specificity, and negative predictive value (NPV) as mean ± standard deviation across the five test folds.

4.2 Fine-Grained MR Severity Grading

Performance metrics reveal that domain-specific foundation models, paired with attentive probes, effectively capture localized hemodynamics. As detailed in Table 4.2, EchoJEPA-L achieved the highest mean F1-score (0.665 ± 0.059). Notably, scaling the frozen backbone to 1 billion parameters with EchoJEPA-G (0.654 ± 0.056) yields highly similar results and nearly identical variance. This indicates that for specialized downstream tasks with limited data, massive backbone scaling offers diminishing returns. Ultimately, EchoJEPA-L yields highly competitive performance on par with the fully supervised 3D CNN ($0.642 \pm$

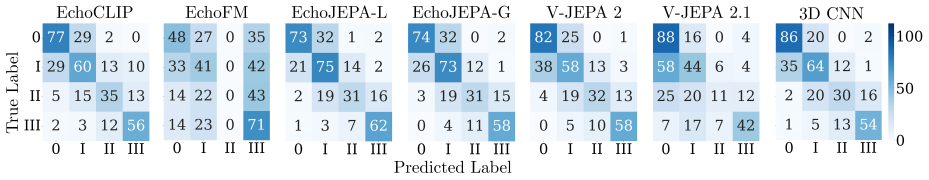

Fig. 2. Confusion matrices detailing the class-wise prediction performance and misclassification distributions for the evaluated foundation models and the 3D CNN baseline across severity grades 0 to III.

0.071), proving that frozen encoders retain rich clinical structures if high-capacity downstream decoders are utilized to route their spatio-temporal features.

Directly comparing JEPA-based architectures provides key insights into the trade-offs between domain-specific and general-purpose pre-training. Although EchoJEPA models outperform their general-purpose counterparts when evaluated with an attentive probe, the general-purpose V-JEPA 2 model remains highly competitive, performing on par with the 3D CNN. Interestingly, when evaluating linear probes, V-JEPA 2 slightly surpasses EchoJEPA. This suggests that general-purpose video models organize features in a more globally linearly separable space, whereas domain-specific models encode dense, localized clinical representations requiring more sophisticated decoding. Furthermore, newer iterations do not strictly guarantee better downstream discrimination; V-JEPA 2.1 unexpectedly underperforms compared to V-JEPA 2, suggesting that while its latent representations may be visually cleaner, they may inadvertently discard the subtle clinical variations necessary for severity grading.

These quantitative insights are supported by the class-wise prediction behaviors illustrated in Figure 2. The confusion matrices demonstrate that EchoJEPA-L maintains a distinct diagonal across all severity grades (0 to III), effectively minimizing both adjacent and severe misclassifications. In contrast, while V-JEPA 2 and the 3D CNN maintain strong overall accuracy, they are more prone to blurring adjacent severity classes, particularly grades I and II. EchoFM struggled significantly across both evaluation frameworks ($F1 = 0.341 \pm 0.039$), indicating that masked autoencoder objectives on echocardiograms may simply fail to capture the localized dynamics of MR.

4.3 Clinically Significant MR

Transitioning to the binary clinical significance task (None/Mild vs. Moderate/Severe), the advantage of domain-specific pre-training diminishes. Predictive capabilities plateau across top architectures (Table 4.3). The fully supervised 3D CNN baseline achieves the highest F1-score (0.902 ± 0.033), outperforming the leading foundation models. Among these, V-JEPA 2 emerges as the closest competitor with an F1-score of 0.877 ± 0.019. Although EchoJEPA-L yields a slightly lower F1-score (0.869 ± 0.020), it achieves a superior ROC-AUC (0.955 ± 0.027), highlighting excellent discriminative power across varying decision thresholds.

Table 3. Performance metrics for 2-class clinical significance classification. Parameters represent probe only. Best results per metric are in **bold**, second best underlined.

	Model	Params	F1	ROC-AUC	Specificity	Sensitivity	Precision	NPV
	3D CNN	30M	**.902** ± .033	.950 ± .024	**.879** ± .033	**.903** ± .033	**.904** ± .035	**.907** ± .046
Attentive	EchoCLIP	2.9M	.853 ± .031	.922 ± .032	.848 ± .011	.853 ± .032	.864 ± .021	.845 ± .050
	EchoFM	11.5M	.649 ± .046	.695 ± .052	.642 ± .048	.649 ± .045	.649 ± .046	.643 ± .047
	EchoJEPA-L	11.5M	.869 ± .020	**.955** ± .027	.853 ± .011	.870 ± .021	.873 ± .021	.864 ± .041
	EchoJEPA-G	21.8M	.876 ± .034	.951 ± .019	.845 ± .028	.878 ± .035	.884 ± .035	.889 ± .056
	V-JEPA 2	21.8M	.877 ± .019	.944 ± .024	.856 ± .017	.878 ± .020	.879 ± .022	.875 ± .032
	V-JEPA 2.1	21.8M	.779 ± .035	.873 ± .037	.718 ± .055	.789 ± .031	.802 ± .040	.813 ± .057

Ultimately, the overlapping cross-validation variance across these leading models suggests that the macroscopic features of clinically significant MR are sufficiently pronounced to be resolved effectively without complex pre-training. Consistent with the four-class findings, EchoFM remains a clear outlier (F1=0.649±0.046), underscoring its persistent limitations even on a simpler binary task, despite sharing an identical parameter footprint with EchoJEPA-L.

4.4 Qualitative Analysis

To qualitatively assess the discriminative power of the evaluated vision foundation models, we visualized their learned feature embeddings using 2D UMAP [12] projections depicted in Figure 3. Each feature vector is first mean-pooled across the sequence dimension to obtain global video representations before applying the downprojection. Both EchoJEPA variants demonstrate distinct class separation. Their projections show regional grouping of the extreme classes, with None (0) and Severe (III) cases localizing to different areas of the embedding space, despite the expected overlap among adjacent intermediate grades. Likewise, V-JEPA 2 successfully separates the extremes but with slightly more boundary mixing in the center compared to EchoJEPA. In contrast, V-JEPA 2.1 displays a noticeably less cohesive structure with fragmented target classes. The lower-performing models exhibit largely entangled latent spaces. EchoCLIP produces scattered embeddings with substantial class overlap, while EchoFM shows little structure, with severity classes distributed nearly uniformly across the manifold.

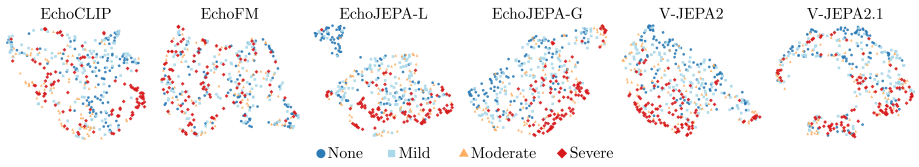

Fig. 3. 2D UMAP visualization of raw features from frozen video foundation models, temporally averaged across the sequence.

5 Conclusion

In this work, we presented a comparative analysis of six spatio-temporal foundation models for mitral regurgitation (MR) grading on a cohort of 413 apical four-chamber Doppler echocardiographic videos. In the four-class severity setting, EchoJEPA-L with an attentive probe achieved the best performance with an F1 score of 0.665, while in the binary clinical significance task, it achieved a superior ROC-AUC (0.955), closely rivaling the fully supervised CNN baseline. Despite the relatively small dataset size, the results are consistent across tasks and demonstrate the effectiveness of self-supervised video representations for echocardiographic MR assessment.

Acknowledgments. The authors acknowledge the financial support by the Faculty of Medicine of the Johannes Kepler University Linz through its internal funding programme SeedplusMED PTM Grant Nr.: SPM-PTM-2025-04. JF and EK acknowledge support from the Austrian Science Fund (FWF) project number 10.55776/COE12.

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

Reproducibility Statement All code necessary to reproduce our results is available at https://github.com/aimi-jku/mr-foundation-assessment.

References

1. Ansari, M.U., Barrios, J.P., Tastet, L., Jhawar, R., Cristin, L., Rich, A., Bibby, D., Fang, Q., Arya, F., Crudo, V., Nguyen, T., Shah, D.J., Delling, F.N., Tison, G.H.: Automated echocardiographic detection of mitral valve prolapse and mitral regurgitation with video-based artificial intelligence algorithms. European Heart Journal - Digital Health **7**(5), ztag061 (06 2026). https://doi.org/10.1093/ehjdh/ztag061, https://doi.org/10.1093/ehjdh/ztag061
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine **30**, 1481–1488 (2024). https://doi.org/10.1038/s41591-024-02959-y
4. Enriquez-Sarano, M., Akins, C.W., Vahanian, A.: Mitral regurgitation. The Lancet **373**(9672), 1382–1394 (Apr 2009). https://doi.org/10.1016/S0140-6736(09)60692-9, https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(09)60692-9/fulltext
5. Figlioli, G., Sticchi, A., Christodoulou, M.N., Hadjidemetriou, A., Amorim Moreira Alves, G., De Carlo, M., Praz, F., Caterina, R.D., Nikolopoulos, G.K., Bonovas, S., Piovani, D.: Global Prevalence of Mitral Regurgitation: A Systematic Review and Meta-Analysis of Population-Based Studies. Journal of Clinical Medicine **14**(8), 2749 (Jan 2025). https://doi.org/10.3390/jcm14082749, https://www.mdpi.com/2077-0383/14/8/2749

6. Holte, E., Podlesnikar, T., Carvalho, F., Demirkiran, A., Manka, R., Martínez, G.G., Michalski, B., Pasquet, A., Separovic Hanzevacki, J., Soliman-Aboumarie, H., Shruti, J.S., Haugaa, K.H., Dweck, M.R.: European Association of Cardiovascular Imaging survey on the evaluation of mitral regurgitation. European Heart Journal - Cardiovascular Imaging **25**(5), 573–578 (May 2024). https://doi.org/10.1093/ehjci/jeae053, https://doi.org/10.1093/ehjci/jeae053
7. Kim, S., Jin, P., Song, S., Chen, C., Li, Y., Ren, H., Li, X., Liu, T., Li, Q.: Echofm: Foundation model for generalizable echocardiogram analysis. IEEE Transactions on Medical Imaging **44**(10), 4049–4062 (2025). https://doi.org/10.1109/TMI.2025.3580713
8. Lin, W., Li, Y., Chen, Y., Hsiao, T., Li, W., Tsai, W., Liu, P., Huang, M.: Automated classification of mitral and tricuspid regurgitation with explainability and real-world practice experience. Journal of the American Heart Association **15**(6), e045851 (2026). https://doi.org/10.1161/JAHA.125.045851, https://www.ahajournals.org/doi/abs/10.1161/JAHA.125.045851
9. Liu, G., Wang, Y., Cheng, H., Shi, Z., Qi, Z., Yao, J., Luo, S., Chen, G.: Automatic segmentation and evaluation of mitral regurgitation using doppler echocardiographic images. Bioengineering **11**, 1131 (11 2024). https://doi.org/10.3390/bioengineering11111131
10. Long, A., Haggerty, C.M., Finer, J., Hartzel, D., Jing, L., Keivani, A., Kelsey, C., Rocha, D., Ruhl, J., vanMaanen, D., Metser, G., Duffy, E., Mawson, T., Maurer, M., Einstein, A.J., Beecy, A., Kumaraiah, D., Homma, S., Liu, Q., Agarwal, V., Lebehn, M., Leon, M., Hahn, R., Elias, P., Poterucha, T.J.: Deep learning for echo analysis, tracking, and evaluation of mitral regurgitation (delineate-mr). Circulation **150**(12), 911–922 (2024). https://doi.org/10.1161/CIRCULATIONAHA.124.068996, https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.124.068996
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101
12. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29), 861 (2018). https://doi.org/10.21105/joss.00861, https://doi.org/10.21105/joss.00861
13. Munim, A., Fallahpour, A., Szasz, T., Attarpour, A., Jiang, R., Sooriyakanthan, B., Sooriyakanthan, M., Whitney, H., Slivnick, J., Rubin, B., Tsang, W., Wang, B.: Echojepa: A latent predictive foundation model for echocardiography. arXiv preprint arXiv:2602.02603 (2026)
14. Mur-Labadia, L., Muckley, M., Bar, A., Assran, M., Sinha, K., Rabbat, M., LeCun, Y., Ballas, N., Bardes, A.: V-jepa 2.1: Unlocking dense features in video self-supervised learning. arXiv preprint arXiv:2603.14482 (2026)
15. Nishimura, R.A., Vahanian, A., Eleid, M.F., Mack, M.J.: Mitral valve disease—current management and future challenges. The Lancet **387**(10025), 1324–1334 (Mar 2016). https://doi.org/10.1016/S0140-6736(16)00558-4, https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)00558-4/fulltext
16. Sadeghpour, A., Jiang, Z., Hummel, Y.M., Frost, M., Lam, C.S., Shah, S.J., Lund, L.H., Stone, G.W., Swaminathan, M., Weissman, N.J., Asch, F.M.: An automated machine learning–based quantitative multiparametric approach for mitral regurgitation severity grading. JACC: Cardiovascular Imaging **18**(1), 1–12 (2025). https://doi.org/https://doi.org/10.1016/j.jcmg.2024.06.011, https://www.sciencedirect.com/science/article/pii/S1936878X2400247X

17. Stone, G.W., Adams, D.H., Abraham, W.T., Kappetein, A.P., Généreux, P., Vranckx, P., Mehran, R., Kuck, K.H., Leon, M.B., Piazza, N., Head, S.J., Filippatos, G., Vahanian, Alec S., f.t.M.V.A.R.C.M.: Clinical trial design principles and endpoint definitions for transcatheter mitral valve repair and replacement: part 2: endpoint definitions: A consensus document from the mitral valve academic research consortium. European Heart Journal **36**(29), 1878–1891 (08 2015). https://doi.org/10.1093/eurheartj/ehv333, https://doi.org/10.1093/eurheartj/ehv333
18. Vrudhula, A., Duffy, G., Vukadinovic, M., Liang, D., Cheng, S., Ouyang, D.: High-throughput deep learning detection of mitral regurgitation. Circulation **150**(12), 923–933 (2024). https://doi.org/10.1161/CIRCULATIONAHA.124.069047, https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.124.069047
19. Zoghbi, W.A., Adams, D., Bonow, R.O., Enriquez-Sarano, M., Foster, E., Grayburn, P.A., Hahn, R.T., Han, Y., Hung, J., Lang, R.M., Little, S.H., Shah, D.J., Shernan, S., Thavendiranathan, P., Thomas, J.D., Weissman, N.J.: Recommendations for Noninvasive Evaluation of Native Valvular Regurgitation: A Report from the American Society of Echocardiography Developed in Collaboration with the Society for Cardiovascular Magnetic Resonance. Journal of the American Society of Echocardiography: Official Publication of the American Society of Echocardiography **30**(4), 303–371 (Apr 2017). https://doi.org/10.1016/j.echo.2017.01.007