HEART-JEPA: LEARNING CROSS-VIEW REPRESENTATIONS FOR CARDIAC MRI WITH JOINT-EMBEDDING PREDICTIVE ARCHITECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Cine cardiac magnetic resonance imaging (CMR) captures the same heart from complementary anatomical views. Learning representations that transfer between these views is a central challenge for reusable cardiac prediction. We introduce HEART-JEPA, a cross-view joint-embedding predictive architecture that predicts masked latent features across anatomical planes without spatially matched patches. Pretrained on 39,992 UK Biobank participants, HEART-JEPA achieves a cross-view structural-transfer R^2 of 0.48 in 15,935 participants absent from pretraining, without target-view refitting or recalibration. Matched ablation studies show substantial gains over same-view pretraining and stronger cardiac phenotype prediction and structural and functional transfer than cross-view contrastive learning. In a separate ventricular-function benchmark, supervised linear probes fitted to concatenated frozen features from all three views achieve the lowest mean LVEF MAE among the seven evaluated models (3.19 percentage points), with competitive RVEF performance. These findings establish cross-view latent prediction as an effective approach to cardiac representation learning and motivate its extension to videos acquired from complementary, spatially unaligned views.

1 INTRODUCTION

Cardiac magnetic resonance imaging (CMR) is widely used for the non-invasive assessment of cardiac anatomy, function, and tissue characteristics (Kramer et al., 2020). Cine CMR captures cardiac motion throughout the cardiac cycle, supporting quantitative assessment of ventricular volumes, myocardial mass, and ejection fraction. Acquisitions along different anatomical planes provide complementary observations of the same heart. Automated assessment of these measurements and other cardiac tasks typically relies on substantial expert annotation and task-specific training (Fu et al., 2026; Shad et al., 2026). Reusable CMR representations offer a common basis for these tasks with fewer task-specific labels.

Self-supervised pretraining and multimodal learning address this annotation burden. CineMA jointly encodes multiple cardiac views and reconstructs masked image patches to learn transferable representations (Fu et al., 2026). ViTa combines multiview masked reconstruction with image–tabular alignment (Zhang et al., 2025), while CMR Transformer learns from correspondence between cine videos and clinical reports (Shad et al., 2026). More recently, MR-JEPA extends LeJEPA (Balestriero & LeCun, 2025) to spatiotemporal CMR inputs and integrates anatomical views through gated attention for downstream prediction (Jacob et al., 2026). These approaches learn cardiac representations through image reconstruction, clinical context, or consistency between augmented observations.

Cross-plane masked latent prediction addresses a different learning problem from joint reconstruction and downstream feature fusion: predicting one anatomical plane from another. This distinction is central to cine CMR. The 2-, 3-, and 4-chamber long-axis views share aspects of cardiac morphology and motion, yet differ in anatomical coverage and spatial layout. The planes are spatially unaligned, so equal image coordinates refer to different anatomy. Paired planes therefore provide complementary context and prediction targets in feature space. We investigate how this prediction task shapes cardiac representations and their transfer across anatomical views.

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.

We introduce HEART-JEPA, a cardiac-specific joint-embedding predictive architecture that extends masked latent prediction from within-view learning to prediction across anatomical views in cine CMR. Given paired 2-, 3-, and 4-chamber acquisitions, HEART-JEPA predicts masked latent features of one cardiac plane from another without requiring patch correspondences or physiological frame alignment. This cross-view objective targets cardiac information that is predictable across complementary anatomical views. Using 39,992 UK Biobank participants, matched experiments show substantial gains in direct transfer without refitting or recalibration. These gains persist in 15,935 participants absent from pretraining, alongside strong cardiac phenotype and ventricular-function prediction. Cross-view latent prediction thus provides an effective inductive bias for reusable representations from multi-view cardiac MRI.

Contributions. (i) Cross-view representation learning. We introduce HEART-JEPA, a cardiac-specific cross-view predictive framework that learns from paired anatomical views by predicting masked latent features across cardiac planes, without requiring supplied patch correspondences or physiological frame alignment. **(ii) Transferable representations across different views.** We demonstrate that cross-view pretraining enables cardiac predictive models trained on one anatomical view to be applied directly to another view without refitting or recalibration. **(iii) Insights into cross-view representation learning.** Matched target-mixture experiments and objective controls reveal distinct trade-offs between linear probing, participant retrieval, and direct transfer. Combining same-view and cross-view prediction produces non-monotonic changes in representation quality, identifying target selection and direct-transfer evaluation as central components of cross-view representation learning.

2 Related Work

Masked and cross-view representation learning. Masked representation learning methods such as MAE, VideoMAE, I-JEPA, and V-JEPA learn visual features by reconstructing or predicting masked content, typically within the same image or video (He et al., 2022; Tong et al., 2022; Assran et al., 2023; Bardes et al., 2024). Multiview methods instead exploit paired observations through contrastive alignment or cross-view completion (Tian et al., 2020; Weinzaepfel et al., 2022). HEART-JEPA combines these ideas in a setting where cardiac views are paired but not spatially aligned: it learns cross-view representations by predicting masked latent features of one anatomical plane from another, without requiring patch correspondence or physiological frame alignment.

Representation learning for cardiac imaging. Recent cardiac representation-learning methods use masked reconstruction, multimodal supervision, or multiview aggregation. CineMA learns from multiview cine CMR through masked reconstruction, ViTa combines imaging with tabular information, CMR Transformer uses image–report supervision, and MR-JEPA applies joint-embedding learning to spatiotemporal CMR (Fu et al., 2026; Zhang et al., 2025; Shad et al., 2026; Jacob et al., 2026). HEART-JEPA focuses on explicit latent prediction across anatomical planes and tests whether the learned representations support direct transfer between cardiac views without target-view refitting or recalibration.

3 Cross-View Latent Prediction

3.1 Cross-View Transfer Without Refitting

For participant n, let x_n^v denote the CMR image from view v and z_n^v its frozen EMA-encoder representation. We first train a cardiac prediction model f_c on features from a source view c, together with a feature standardizer S_c. We then apply the same model directly to features from a different target view t:

$$\hat{y}_{n,c\to t} = f_c\big(S_c\big(z_n^t\big)\big), \qquad c \neq t. \tag{1}$$

This protocol tests direct transfer using encoder representations alone. The source-fitted standardizer and prediction model remain fixed, and the pretraining predictor is excluded from evaluation.

Direct transfer reflects the combined effects of view-specific feature means, scales, directions, and anatomical coverage. A negative R^2 quantifies poor transfer rather than identifying a single source of mismatch. The 2Ch and 3Ch views predominantly capture the left heart, whereas 4Ch includes both

Table 1: **Study cohorts and evaluation roles.**

Label	Sample size	Definition	Evaluation role
C1	39,992	Pretraining cohort	Self-supervised pretraining and monitoring
C2	17,841	Participant-disjoint from C1	Main held-out evaluation
C3	15,935	Subset of C2 with available cardiac IDPs	Cardiac probing, EF, and disease evaluation
C4	4,096	Fixed subset of C3	Centered representation geometry and linear CKA
C5	512	Fixed subset of C3	Temporal-token matching across views
C6	360	External M&Ms-2 cohort; 4Ch only	External cardiac classification

left- and right-sided chambers (Kramer et al., 2020). Right-heart prediction from left-heart views can draw on indirect anatomical correlates. Appendix I (Table S10) reports separate LV/LA and RV/RA summaries.

3.2 Masked prediction across anatomical planes

Each sample provides paired 2-, 3-, and 4-chamber cine CMR videos from the same participant. Each input contains 50 frames at 128×128 resolution and is tokenized using a temporal tube size of 5 and a spatial patch size of 16, yielding the spatiotemporal lattice $\mathcal{G} = \{1, \dots, 10\} \times \{1, \dots, 8\} \times \{1, \dots, 8\}$.

For an ordered context–target view pair (v_c, v_t), a context mask $M_c \subset \mathcal{G}$ removes tokens from the online encoder. Positional embeddings are added before masking, and only tokens in $\mathcal{G} \setminus M_c$ are encoded. The EMA target encoder, in contrast, processes the complete target-view video. Writing $h^c = E_\theta(x^{v_c}; \mathcal{G} \setminus M_c)$ and $H^t = E_\phi(x^{v_t}; \mathcal{G})$, the cross-view latent prediction loss is

$$\mathcal{L}_{\text{token}} = \frac{1}{|Q_t|} \sum_{q \in Q_t} \left\| P_\psi(h^c, q, v_t) - \text{sg}(H^t_q) \right\|_2^2 . \tag{2}$$

For each target query q, the predictor receives the visible context features, a factorized target-position embedding, and the target-view identity. The stop-gradient operator sg prevents gradient updates through the target branch, while the target encoder is updated by exponential moving average, $\phi \leftarrow \tau\phi + (1-\tau)\theta$, with $\tau = 0.996$. The encoder contains 12 Transformer blocks and the predictor contains six (Appendix B).

HEART-JEPA uses distinct anatomical planes for context and target, $v_c \neq v_t$. In the primary configuration, $Q_t = M_c$: context-mask indices also specify queried target positions. These indices identify positions on separately normalized token grids, not registered anatomy. Training uses paired videos without cross-view image registration, patch matching, or physiological frame alignment. Independently sampled target queries retain positive transfer (Section 5.2; Appendix D). Under the fixed-target assumptions in Appendix A, the optimal predictor averages out target-specific variation that is unpredictable from context. This result motivates the prediction objective; compatibility of encoder features across views is a separate empirical question.

4 Experimental Design

Data and preprocessing. Pretraining uses 39,992 UK Biobank participants. Each study includes all three long-axis views (Bycroft et al., 2018); 37,993 studies are used for optimization and 1,999 for monitoring. All controlled models share segmentation-derived bounding-box crops, per-frame z-score normalization, and resizing. Segmentation masks define the crops rather than encoder inputs, supplying a shared anatomical prior. Inputs retain the first 50 frames or repeat the last frame when shorter, without temporal interpolation or physiological registration.

Cohorts and separation. Table 1 summarizes the cohorts used throughout the experiments. The full mixture sweep and objective-control probing and transfer analyses use C2; the participant-disjoint cross-view versus same-view endpoint comparison uses C3 (Appendix C).

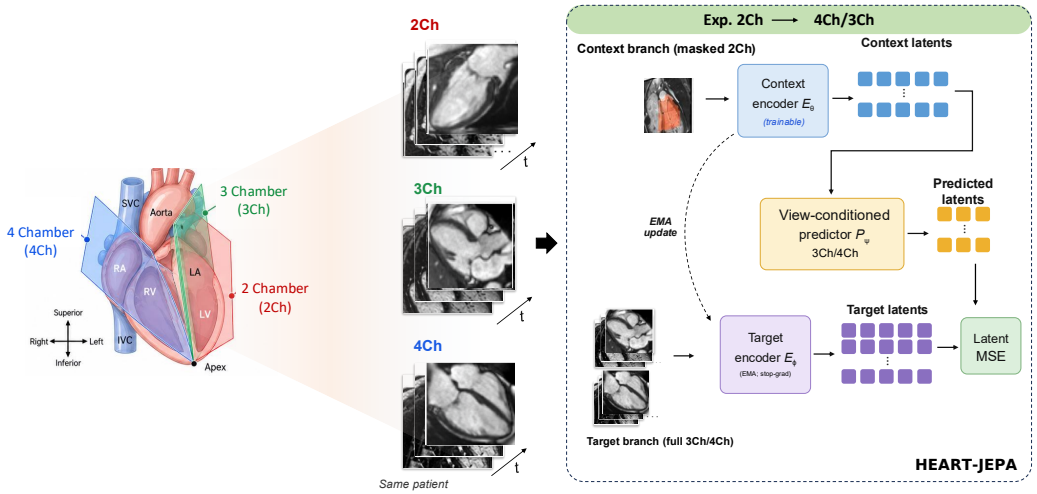


Figure 1: **Overview of HEART-JEPA.** Paired 2-, 3-, and 4-chamber cine CMR views from the same subject provide complementary observations of the same heart. The right panel illustrates a 2Ch→3Ch/4Ch training configuration: a masked 2Ch context view is encoded by the trainable context encoder E_θ, and a view-conditioned predictor P_ψ predicts latent features at queried positions in the full 3Ch or 4Ch target view. Target features are produced by the EMA-updated encoder E_ϕ with stop-gradient, and training minimizes the latent-space mean-squared error between predicted and target features.

Controlled comparisons. We vary the probability of a same-view target, $p_{\text{same}} \in \{0, 0.25, 0.5, 0.75, 1\}$, using three matched training seeds. Within each seed, only p_{same} varies: all settings share a 12-block ViT encoder, a six-block predictor, MultiBlockMask3D masking, and identical data and preprocessing. The ordered study split is fixed across all runs. Training uses AdamW with batch size 48 for 276,850 updates, an EMA coefficient of 0.996, and the fixed epoch-350 checkpoint without early stopping. We additionally evaluate independent target queries, Global HEART-JEPA, and cross-view EMA-InfoNCE. Global HEART-JEPA is the global non-contrastive variant of HEART-JEPA, using normalized regression on pooled video representations in place of token-level MSE. The comparison evaluates objective designs that differ in prediction granularity, projection, and normalization. Appendices B–C specify the configurations and matched components.

Probing and cross-view transfer. For cardiac probing, we concatenate the frozen 2Ch, 3Ch, and 4Ch representations and fit standardized linear models (probes) to 17 imaging-derived phenotypes (IDPs), reporting five-fold cross-validated R^2. For cross-view transfer, a standardized ridge model ($\alpha = 1$) is trained on one source view and applied unchanged to a different target view without refitting or recalibration (Equation (1)). Performance is averaged across five participant folds and six ordered view pairs. The 13 structural and four functional IDP groups are defined in Appendix I.

Ventricular function prediction. We evaluate frozen representations for LVEF and RVEF prediction on the 15,935 participants in C3. Concatenated 2Ch/3Ch/4Ch features are used to train supervised linear probes for seven representation models using the same participants, probe specification, and outer test folds. We report MAE in EF percentage points and Pearson correlation R; full cross-validation and aggregation details are provided in Appendix I.1.

Retrieval, geometry, and temporal matching. We further characterize cross-view representations through participant retrieval and feature-space geometry. Retrieval uses cosine similarity without a learned metric. Centered cosine similarity and linear CKA are evaluated on C4, while C5 is used to

measure cross-view temporal-token matching at normalized sequence positions, using acquisition indices rather than physiologically registered cardiac phases.

Additional utility benchmarks. We evaluate representation utility under different available-view and limited-label settings and assess external transfer using frozen 4Ch features on M&Ms-2 (Martín-Isla et al., 2021). Comparisons include HEART-JEPA, CMR Transformer, CineMA, ViTa-MAE, and matched internal controls. All evaluation splits are participant-level; additional details on checkpoints, aggregation, and contextual comparators are provided in the Appendix.

5 RESULTS

5.1 CROSS-VIEW PREDICTION ENABLES DIRECT TRANSFER ACROSS CARDIAC VIEWS

Cross-view prediction substantially improves direct transfer between anatomical views. In the matched five-condition target-mixture experiment on C2, pure cross-view pretraining achieves the highest mean structural-transfer R^2 (Figure 2; Table S2). Compared with pure same-view pretraining, structural-transfer R^2 increases from -1.039 ± 0.393 to 0.479 ± 0.022. Here, a prediction model trained on one cardiac view is applied unchanged to another view without target-view refitting or recalibration. Cross-view pretraining also improves fitted three-view cardiac probing, from 0.677 ± 0.021 to 0.743 ± 0.007, although the change is substantially larger for direct cross-view transfer.

The effect persists after excluding all pretraining participants. On C3, which excludes all participants present in pretraining, cross-view pretraining maintains a structural-transfer R^2 of 0.477 ± 0.020, compared with -1.101 ± 0.393 for same-view pretraining. Participant retrieval shows a similarly large difference, with top-1 accuracy of $41.24 \pm 5.43\%$ versus $0.0279 \pm 0.0103\%$. On C4, centered matched cosine is 0.787 ± 0.027 for cross-view pretraining versus 0.046 ± 0.021 for same-view pretraining, while linear CKA is 0.850 ± 0.007 versus 0.055 ± 0.030, respectively (Figure 3b–c; Table S7). Together, these results show that cross-view prediction changes both the transferability and the geometry of the learned representations.

Transferable representations retain view identity. View identity remains readily accessible in the cross-view representations. Linear view classification reaches $99.98 \pm 0.0024\%$ for cross-view pretraining and $99.98 \pm 0.0175\%$ for same-view pretraining, with all target-mixture means exceeding 99.97%. The representations therefore retain view identity alongside the cardiac information used for direct transfer across anatomical planes. Trait-level analyses show stronger transfer for left-ventricular wall thickness, myocardial mass, and left-atrial volumes than for right-atrial volumes (Appendix I).

5.2 PROBING, RETRIEVAL, AND CROSS-VIEW TRANSFER CAPTURE DISTINCT PROPERTIES

Fitted prediction and direct transfer favor different objectives. Global HEART-JEPA achieves a higher fitted three-view cardiac-probe R^2 than HEART-JEPA (0.775 ± 0.002 versus 0.743 ± 0.007), yet its single-view structural-transfer R^2 is negative (-0.550 ± 0.130). Its high raw cosine similarity is also accompanied by a small matched–mismatched margin (0.008 ± 0.002). These results separate the ability to fit a cardiac prediction model from the ability to transfer it unchanged across views (Figure 3a).

Retrieval and cardiac transfer favor different objectives. The contrastive control achieves higher mean C3-query retrieval than HEART-JEPA ($45.67 \pm 1.27\%$ versus $41.24 \pm 5.43\%$) and a larger cosine margin in all three paired seeds. In contrast, HEART-JEPA achieves higher fitted cardiac probing (0.743 ± 0.007 versus 0.725 ± 0.002) and structural transfer (0.479 ± 0.022 versus 0.442 ± 0.023). Functional transfer shows the same ordering in all three paired seeds, whereas retrieval favors contrastive learning in two of three (Appendix E). These results show that participant retrieval and transferable cardiac representations capture related but distinct properties.

Independent target queries preserve cross-view transfer. Using independently sampled target queries retains a structural-transfer R^2 of 0.470 ± 0.027, compared with 0.479 ± 0.022 for shared

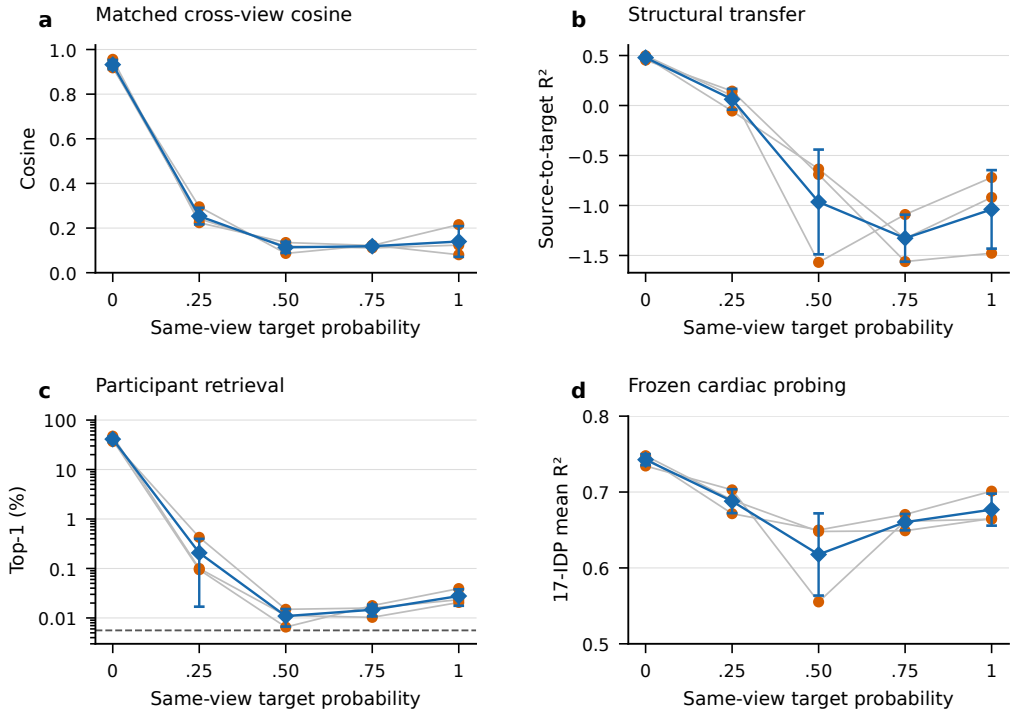


Figure 2: **Effect of target source on learned representations.** (a) Matched cross-view cosine similarity, (b) structural transfer without target-view refitting, (c) participant retrieval, and (d) fitted three-view cardiac probing. Points denote individual training seeds, gray lines connect matched seeds, and diamonds show mean ± sample SD. Retrieval is shown on a logarithmic percentage scale, with the dashed line indicating random top-1 performance. All conditions are trained from scratch under the same training budget.

queries. Fitted cardiac-probe performance is similar, while functional transfer favors independent queries in all three seeds (Appendix D). Independent sampling preserves incidental mask overlap (Jaccard 0.682 ± 0.072) while removing exact index coupling. Positive transfer under this intervention establishes robustness to that coupling choice.

5.3 Target mixtures reveal nonlinear objective behavior

Target mixtures produce non-monotonic changes in representation quality. Introducing 25% same-view targets reduces all four mean evaluation metrics relative to pure cross-view prediction (Figure 2; Table S2), although fitted cardiac probing and structural transfer remain above the pure same-view endpoint. At $p_{\text{same}} = .50$, fitted cardiac probing decreases to 0.618 ± 0.054, below both pure endpoints, whereas structural transfer reaches its lowest mean at $p_{\text{same}} = .75$. Thus, changing the target-view distribution alters the learned representation nonlinearly rather than interpolating between independently learned same-view and cross-view solutions.

Final-checkpoint gradient diagnostics are consistent with this change. The cosine between same-view and cross-view gradients in the final encoder block is 0.469 ± 0.078 at $p_{\text{same}} = 0$ and 0.026 ± 0.005 at .25 (Appendix H). These measurements describe local gradient geometry at the final checkpoint; the training mechanism remains unresolved.

5.4 Cross-view structure extends to temporal representations

Cross-view organization is also evident below the pooled video representation. On C5, we match spatially averaged temporal tubelets across anatomical planes within each participant. Sequence-position-centered top-1 retrieval reaches $68.49 \pm 3.56\%$ for cross-view pretraining, compared with

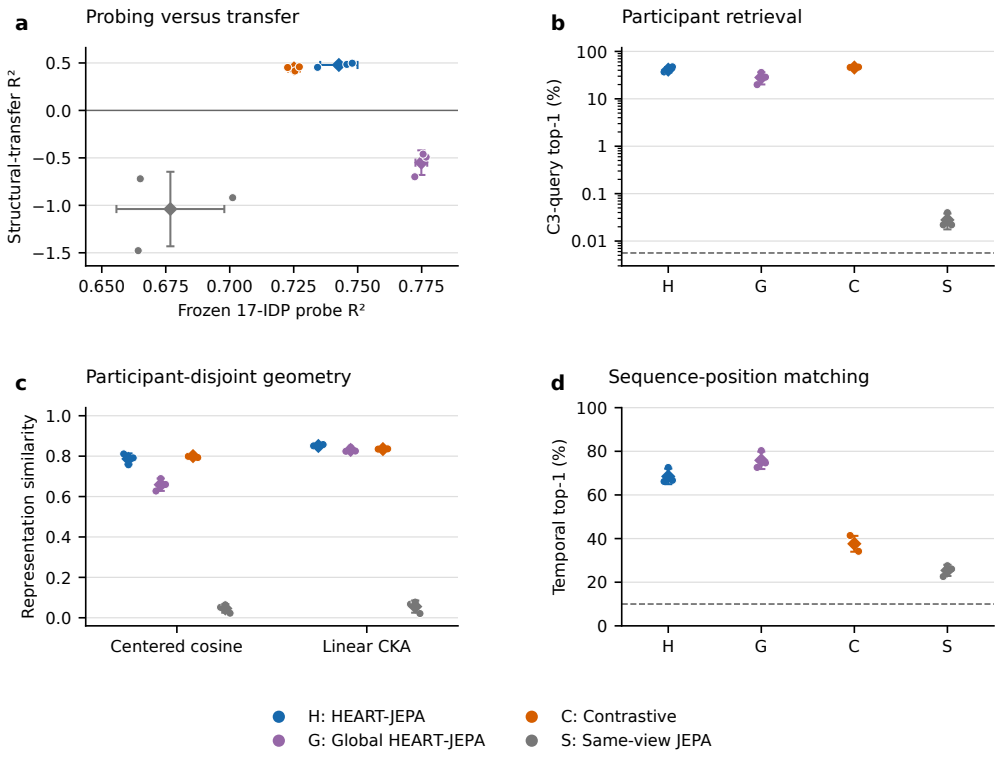


Figure 3: **Different evaluation protocols capture distinct representation properties.** (a) Three-view fitted cardiac probing versus single-view structural transfer on C2; bars show SD in both coordinates. (b) Participant retrieval using C3 queries and the C2 gallery. (c) Centered cross-view representation geometry on C4. (d) Sequence-position-centered temporal matching on C5. Dots denote training seeds; diamonds and bars show mean ± sample SD. Dashed lines indicate random baselines where applicable. High geometric similarity or retrieval performance does not by itself imply strong direct cross-view transfer.

$25.41 \pm 2.55\%$ for same-view pretraining and a 10% chance level. Normalized circular index error is correspondingly lower, at 0.061 ± 0.011 versus 0.196 ± 0.008 (Figure 3d; Appendix G). Thus, the cross-view structure learned by HEART-JEPA extends to temporal token representations. The metric measures correspondence between normalized acquisition indices rather than physiologically synchronized cardiac phases.

5.5 Frozen representations support ventricular function prediction

Frozen HEART-JEPA representations also support ventricular function prediction. Among the seven evaluated models, HEART-JEPA achieves the lowest mean LVEF MAE (3.1888 percentage points) and the highest LVEF Pearson R (0.7563; Table 2). Relative to cross-view contrastive learning, its LVEF and RVEF MAEs are lower by 6.7% and 4.2%, respectively. HEART-JEPA also improves both ventricular endpoints over same-view JEPA and the contextual pretrained comparators.

Global HEART-JEPA achieves the strongest RVEF results, with MAE 3.2307 compared with 3.3211 for HEART-JEPA, and Pearson R of 0.7267 compared with 0.7092. Its LVEF MAE is also close (3.2016 versus 3.1888). Together, these results show that cross-view latent prediction yields strong frozen cardiac representations with task-dependent strengths.

The external systems benchmark complete pretrained representations under a common downstream protocol. MR-JEPA's published EF results use a different cohort and input protocol (Jacob et al., 2026) and are therefore outside this comparison.

Table 2: Frozen-feature LVEF and RVEF prediction on the same 15,935 participants (C3), three-view inputs, and outer test folds. Values are mean ± sample SD over three five-fold CV repetitions. Metrics use pooled out-of-fold predictions, with checkpoint metrics averaged within each repetition. SD measures split sensitivity conditional on the evaluated checkpoints. MAE is in EF percentage points; R is Pearson correlation. External models differ in pretraining. Bold marks the best mean; Appendix I.1 gives the aggregation.

		LVEF		RVEF	
Model	Pretrained checkpoints	MAE ↓	R ↑	MAE ↓	R ↑
HEART-JEPA	3	$\mathbf{3.1888} \pm 0.0022$	$\mathbf{0.7563} \pm 0.0005$	3.3211 ± 0.0017	0.7092 ± 0.0004
Global HEART-JEPA	3	3.2016 ± 0.0014	0.7527 ± 0.0003	$\mathbf{3.2307} \pm 0.0023$	$\mathbf{0.7267} \pm 0.0003$
Same-view JEPA	3	3.4223 ± 0.0018	0.7085 ± 0.0006	3.5126 ± 0.0008	0.6689 ± 0.0003
Cross-view contrastive	3	3.4180 ± 0.0017	0.7084 ± 0.0005	3.4674 ± 0.0035	0.6791 ± 0.0008
CMR Transformer	1	3.2390 ± 0.0029	0.7451 ± 0.0004	3.5207 ± 0.0029	0.6696 ± 0.0010
CineMA	1	4.0860 ± 0.0031	0.5254 ± 0.0011	3.9496 ± 0.0043	0.5427 ± 0.0014
ViTa-MAE	1	3.8391 ± 0.0124	0.6043 ± 0.0029	3.9735 ± 0.0050	0.5530 ± 0.0017

5.6 Cross-view representations generalize across cardiac tasks and settings

The utility of the learned representations extends beyond the primary IDP and ventricular-function analyses. On C3, HEART-JEPA achieves higher mean ROC-AUC and average precision than same-view JEPA across all five UKB disease-label endpoints. Relative to contrastive learning, it achieves higher ROC-AUC for atrial fibrillation and heart failure and higher average precision on four endpoints, while the remaining comparisons favor contrastive learning (Appendix L). These evaluations measure discrimination of recorded disease labels, with diagnosis timing outside the analysis.

The separate benchmark checkpoint further supports available-view and limited-label evaluations. On C2, mean 17-IDP R^2 is 0.754 ± 0.005 across five probe folds, while two-view probes retain R^2 values of 0.702–0.742. With only 1%, 5%, and 10% of the labeled training pool, mean R^2 reaches 0.587, 0.681, and 0.702, respectively, averaged over five seeded 80/20 splits (Appendices I–J).

External M&Ms-2 evaluation provides a complementary test on a separate cardiac dataset. CMR Transformer achieves higher macro AUC for disease-versus-normal classification (0.898 versus 0.848) and one-versus-rest classification (0.801 versus 0.766), whereas HEART-JEPA has the highest displayed mean for hypertrophic cardiomyopathy versus normal and tetralogy of Fallot versus rest (Appendix K). The external benchmark identifies model-specific strengths in downstream classification.

6 Discussion and Limitations

HEART-JEPA learns representations that support direct cross-view prediction while retaining view-specific information. The controls distinguish three representation capabilities: fitted prediction, participant retrieval, and direct transfer. Global HEART-JEPA leads in fitted phenotype probing and RVEF prediction, contrastive learning favors retrieval, and HEART-JEPA combines strong direct transfer with the lowest mean LVEF error in the separate three-view EF benchmark.

HEART-JEPA learns within-modality relationships between anatomical planes through masked latent prediction with a shared encoder and an EMA target branch. CrossJEPA (Perera et al., 2025) instead learns 3D point-cloud representations by predicting rendered 2D image features from a frozen teacher. The distinction is between cross-view transfer within a modality and cross-modal distillation.

The matched target-mixture experiment isolates target source; the external UK Biobank benchmarks compare complete pretrained systems. Their pretraining data, supervision, and input processing differ. Exact pretraining participant lists were unavailable for the external models, so overlap with C3

remains unverified. C3 establishes participant-disjointness for our own pretraining. These comparison settings support different levels of attribution: target-source effects in the matched experiment and downstream utility in external benchmarks.

The non-monotonic mixture response identifies target selection as a consequential design choice. Competition between the objectives within a shared encoder is a hypothesis for the performance dip; its mechanism remains unresolved. Direct transfer also reflects view-dependent feature means and scales and unequal anatomical coverage. Temporal matching measures sequence-position correspondence, while M&Ms-2 evaluates external downstream classification. Physiological phase alignment and external direct cross-view transfer remain outside these evaluations.

These findings motivate extending video JEPA to paired camera views of the same scene or event. Natural-video evaluation would test whether these transfer properties persist under temporal offsets, occlusion, and unequal scene coverage.

7 Conclusion

HEART-JEPA learns cross-view cardiac representations by predicting target-view latent features from masked context across paired anatomical views. Matched experiments show that cross-view pretraining substantially improves direct transfer of cardiac prediction models across views without target-view refitting or recalibration, while preserving view-specific information. Frozen-feature evaluations further demonstrate strong downstream cardiac utility, including competitive ventricular-function and disease prediction. Together, these results identify cross-view latent prediction as an effective strategy for learning transferable representations from multi-view cine CMR.

AI USE STATEMENT

We used generative AI tools for language editing and manuscript organization; drafting cohort-selection and reproducibility descriptions from author-provided study information; literature identification; clarification of the theoretical argument; and feedback on methodology and result interpretation. They also assisted with incorporating reviewer feedback, checking main-text and appendix consistency, and preparing tables and figure labels from author-supplied summaries and saved aggregate results. The authors are responsible for verifying this content against study records, analysis outputs, and original sources and for the accuracy and integrity of the manuscript.

ETHICS STATEMENT

This study used de-identified UK Biobank and M&Ms-2 data under their respective access requirements. The models are for research use; clinical validation is outside the present study.

UK Biobank data are available to approved researchers and cannot be redistributed by the authors. M&Ms-2 access follows the challenge conditions. The supplementary source project provides aggregate results and provenance metadata and excludes participant identifiers.

REPRODUCIBILITY STATEMENT

Sections 3–4 and Appendices B–C describe the objective, architecture, preprocessing, and controlled training design. The source project provides aggregate results, generated tables, data provenance, and figure assets for inspecting the reported analyses. Full replication additionally requires the training and evaluation implementation and approved access to restricted UK Biobank data.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://doi.org/10.1109/CVPR52729.2023.01499.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. doi: 10.48550/arXiv.2511.08544. URL https://arxiv.org/abs/2511.08544.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018. doi: 10.1038/s41586-018-0579-z. URL https://www.nature.com/articles/s41586-018-0579-z.

Yunguan Fu, Wenjia Bai, Weixi Yi, Charlotte Manisty, Anish N. Bhuva, Thomas A. Treibel, James C. Moon, Matthew J. Clarkson, Rhodri Huw Davies, and Yipeng Hu. Development and validation of a versatile foundation model for cine cardiac magnetic resonance image analysis. *Communications Medicine*, 6(1):408, 2026. doi: 10.1038/s43856-026-01636-0. URL https://www.nature.com/articles/s43856-026-01636-0.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

Athira J. Jacob, Puneet Sharma, Dorin Comaniciu, and Daniel Rueckert. MR-JEPA: A general purpose video foundation model for cardiac MRI. *arXiv preprint arXiv:2608.30975*, 2026. doi: 10.48550/arXiv.2608.30975. URL https://arxiv.org/abs/2608.30975.

Christopher M. Kramer, Jörg Barkhausen, Chiara Bucciarelli-Ducci, Scott D. Flamm, Raymond J. Kim, and Eike Nagel. Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. *Journal of Cardiovascular Magnetic Resonance*, 22(1):17, 2020. doi: 10.1186/s12968-020-00607-1. URL https://doi.org/10.1186/s12968-020-00607-1.

Carlos Martín-Isla, José F. Rodríguez Palomares, Andrea Guala, Sergio Escalera, and Karim Lekadir. Multi-disease, multi-view & multi-center right ventricular segmentation in cardiac MRI (M&Ms-2). STACOM Challenge protocol, Zenodo, 2021. URL https://zenodo.org/records/4573984.

Avishka Perera, Kumal Hewagamage, Saeedha Nazar, Kavishka Abeywardana, Hasitha Gallella, Ranga Rodrigo, and Mohamed Afham. CrossJEPA: Cross-modal joint-embedding predictive architecture for efficient 3D representation learning from 2D images. *arXiv preprint arXiv:2511.18424*, 2025. doi: 10.48550/arXiv.2511.18424.

Rohan Shad, Cyril Zakka, Dhamanpreet Kaur, Mrudang Mathur, Robyn Fong, Joseph Cho, Ross Warren Filice, John Mongan, Kimberly Kallianos, Nishith Khandwala, David Eng, Matthew Leipzig, Walter R. Witschey, Alejandro de Feria, Victor A. Ferrari, Euan A. Ashley, Michael A. Acker, Curtis Langlotz, and William Hiesinger. A generalizable deep learning system for cardiac MRI. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01637-3. URL https://www.nature.com/articles/s41551-026-01637-3.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision – ECCV 2020*, volume 12356 of *Lecture Notes in Computer Science*, pp. 776–794, 2020. doi: 10.1007/978-3-030-58621-8_45. URL https://link.springer.com/chapter/10.1007/978-3-030-58621-8_45.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022. doi: 10.52202/068431-0732. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.

Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. CroCo: Self-supervised pre-training for 3D vision tasks by cross-view completion. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3502–3516, 2022. doi: 10.52202/068431-0253. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/16e71d1a24b98a02c17b1be1f634f979-Abstract-Conference.html.

Yundi Zhang, Paul Hager, Che Liu, Suprosanna Shit, Chen Chen, Daniel Rueckert, and Jiazhen Pan. Towards cardiac MRI foundation models: Comprehensive visual-tabular representations for whole-heart assessment and beyond. *Medical Image Analysis*, 106:103756, 2025. doi: 10.1016/j.media.2025.103756. URL https://www.sciencedirect.com/science/article/pii/S1361841525003032.

A Why Cross-View Prediction Favors Shared Information

The fixed-target squared-loss objective in Equation (2) has a population-optimal predictor that averages out conditionally centered, unpredictable target variation. We derive this property under explicit assumptions and distinguish it from direct transfer of encoder features.

The optimal fixed-target predictor. Fix an ordered view pair and a target position. Let U contain the visible context and the query information supplied to the predictor, and let Y be the corresponding fixed, square-integrable target feature. With $m(U) = \mathbb{E}[Y \mid U]$, any square-integrable predictor f satisfies

$$\mathbb{E}\|Y - f(U)\|_2^2 = \mathbb{E}\|Y - m(U)\|_2^2 + \mathbb{E}\|m(U) - f(U)\|_2^2. \tag{3}$$

To see this, expand $Y - f(U) = (Y - m(U)) + (m(U) - f(U))$. The cross term is zero because $\mathbb{E}[Y - m(U) \mid U] = 0$. Hence $f^*(U) = \mathbb{E}[Y \mid U]$ minimizes the population loss.

Suppose the fixed target can be written as

$$Y = A_t C + B_t S_t + \epsilon_t, \tag{4}$$

where C contains information shared across views, S_t is target-specific variation, and A_t, B_t are fixed linear maps. Assume $S_t \perp U \mid C$, $\mathbb{E}[S_t \mid C] = 0$, and $\mathbb{E}[\epsilon_t \mid U] = 0$. The tower property gives

$$\mathbb{E}[S_t \mid U] = \mathbb{E}[\mathbb{E}[S_t \mid U, C] \mid U] = \mathbb{E}[\mathbb{E}[S_t \mid C] \mid U] = 0. \tag{5}$$

It follows that

$$f^*(U) = A_t \mathbb{E}[C \mid U]. \tag{6}$$

Under these assumptions, the optimal predictor retains the context-predictable part of the shared state and averages out unpredictable target-specific variation. The assumptions apply to each view pair and queried position, including the masking and query-selection rule. Acquisition factors that are predictable across views can be retained.

Encoder compatibility is a separate property. Let two views represent the same scalar by $z^c = s$ and $z^t = -s$. A predictor can translate perfectly between them, but a source-view prediction model $f_c(z) = z$ fails when applied unchanged to z^t. Thus, accurate cross-view latent prediction does not guarantee accurate predictions under the transfer protocol in Equation (1). The derivation holds the target fixed, whereas EMA training updates it jointly with the online encoder. Constant representations and information unused by the predictor remain compatible with the fixed-target analysis. We therefore evaluate encoder compatibility empirically through cardiac prediction, retrieval, geometry, and view classification.

B Implementation Details

Architecture and inputs. The encoder uses 12 Transformer blocks, six attention heads, and 384-dimensional tokens. The predictor uses six Transformer blocks, dimension 128, and four heads. The predictor takes 384-dimensional context tokens and uses a 128-dimensional internal representation. Its output in Equation (2) is required to match the 384-dimensional target tokens. Each input has 50 frames at 128×128 resolution, with spatial patch size 16 and temporal tube size 5, yielding a native $10 \times 8 \times 8$ token grid. MultiBlockMask3D samples eight target blocks with spatial scale $[0.15, 0.45]$, aspect ratio $[0.75, 1.5]$, and temporal scale $[0.5, 1.0]$. The EMA coefficient is 0.996. The target-view identity is supplied only to the predictor. The full model contains approximately 44.9M parameters, of which 23.1M are optimized directly; downstream evaluations freeze the EMA target encoder. Formal runs use 791 optimizer updates per epoch and 276,850 updates in total, with a constant learning rate of 1.5×10^{-4} and no warmup or decay. The same schedule and optimization budget apply to every condition. Video preprocessing retains the first 50 frames or repeats the last frame when fewer are available; it does not temporally interpolate or register cardiac phases. Crops use the union of nonzero segmentation pixels across frames, and each cropped frame is z-scored before bilinear resizing. The ViTa-MAE comparator is its image-only checkpoint; the evaluation adapter repeats a long-axis view across its nine expected spatial slices. This adapter is part of the external-model evaluation protocol. All controlled objectives share the segmentation-derived crop prior. Padding prevalence and sensitivity to temporal resampling remain unmeasured.

Context masks and target queries. Let M_c be the union of eight sampled MultiBlockMask3D regions. Spatial and temporal positional embeddings are added to all tokens before the context sequence is restricted to $\mathcal{G} \setminus M_c$. In the primary model, the predictor queries and the target features selected for the loss use $Q_t = M_c$. These shared indices identify positions on separately normalized $10 \times 8 \times 8$ lattices; the target encoder processes the full video without explicit plane registration. Each update samples one ordered view pair and one M_c for the minibatch. The independent-query control obtains Q_t via a second independent draw from the same masking distribution. Across 10,000 fixed-seed audit draws, the two sets were never identical: they contained 470.97 ± 47.62 and 470.61 ± 47.36 tokens, respectively, with a mean Jaccard overlap of 0.682 ± 0.072. This intervention removes exact set coupling while preserving incidental overlap and the marginal target-count distribution. The cardinalities $|Q_t|$ and $|M_c|$ can differ within an update.

Global HEART-JEPA. Global HEART-JEPA is the global non-contrastive variant of HEART-JEPA. It applies cross-view latent prediction to one pooled video representation instead of a set of target tokens. The online branch mean-pools the visible context tokens from the masked context view and adds a learned embedding specifying the target view. The resulting representation is passed through an online projector and prediction MLP. In the target branch, all tokens from the full target view are mean-pooled and passed through an EMA-updated copy of the projector. The loss is the squared Euclidean distance between the ℓ_2-normalized 128-dimensional online prediction and the stop-gradient target projection, without negative samples. EMA updates the target projector parameters; its BatchNorm layers use target-batch statistics and maintain branch-specific running buffers. The stop-gradient operation excludes the target branch from backpropagation.

To control for head capacity, we set the projector and predictor hidden dimensions to 1,649 and 1,743, respectively, yielding 1,298,688 trainable parameters in the online head, exactly matching the token predictor. The additional EMA projector increases the total parameter count from 44.9M to 45.8M. Predicting one representation per video reduces head-level FLOPs relative to predicting a set of target tokens.

Using the encoder notation in Section 3.2, let $h^c = E_\theta(x^{v_c}; \mathcal{G} \setminus M_c)$ and $H^t = E_\phi(x^{v_t}; \mathcal{G})$. Write $\mathrm{pool}(\cdot)$ for averaging the supplied tokens and $\mathrm{norm}(u) = u/\|u\|_2$. Writing both outputs in the same 128-dimensional projection space, the loss is

$$\mathcal{L}_{\mathrm{global}} = \left\| \mathrm{norm}(q_\psi(g_\psi(\mathrm{pool}(h^c) + e_{v_t}))) - \mathrm{sg}\big[\mathrm{norm}\big(g_\xi\big(\mathrm{pool}(H^t)\big)\big)\big] \right\|_2^2 . \tag{7}$$

Here g_ψ and q_ψ are the online projector and predictor, g_ξ is the EMA target projector, and e_{v_t} is the target-view embedding. Context pooling uses only visible tokens; target pooling uses the full target grid.

Cross-view contrastive learning. The EMA-InfoNCE control instead discriminates participants: the masked context is the query, a full distinct view of the same participant is its positive key, and other participants in the minibatch provide negative keys. It shares the backbone and training protocol but uses a global projection head rather than the token predictor. These objective-family comparisons jointly vary pooling, normalization, head design, and the use of negatives.

C Training Design and Cohort Separation

The target-mixture experiment changes only p_{same} within each training seed. All five settings share the manifest, ordered study split, architecture, masking, preprocessing, optimizer, update count, and checkpoint rule. Seeds 42, 43, and 44 give three stochastic training runs on the same ordered split (split seed 42). Table S1 summarizes what is held constant and what changes in the additional controls.

Table S1: Controlled training design. The formal mixture includes $p_{\text{same}} \in \{0, .25, .5, .75, 1\}$. All columns use the same manifest, ordered split, deterministic preprocessing, batch size 48, AdamW, learning rate 1.5×10^{-4}, weight decay 0.05, constant schedule, EMA coefficient 0.996, 276,850 updates.

	Target mixture	Independent queries	Global HEART-JEPA	Contrastive
Encoder	ViT-12	ViT-12	ViT-12	ViT-12
Prediction unit	Target tokens	Target tokens	Pooled video	Pooled video
Objective	Token MSE	Token MSE	Normalized regression	EMA-InfoNCE
Trainable parameters	23.1M	23.1M	23.1M	22.1M
Total parameters	44.9M	44.9M	45.8M	—
Context mask	MultiBlockMask3D	Same distribution	Same distribution	Same distribution
Target query indices	M_c	Independent Q_t	Not applicable	Not applicable
Training seeds	42, 43, 44	42, 43, 44	42, 43, 44	42, 43, 44

The independent-query control changes only the target-query draw. Global HEART-JEPA matches the number of trainable head parameters, but its EMA projector changes the total parameter count and its single prediction reduces head-level compute. The contrastive head has fewer trainable parameters. These controls compare objective designs with different head-level compute as well as prediction granularity.

Checkpoint scope across evaluations. The EF benchmark is reported separately from the matched intervention. Table 2 lists three checkpoints each for HEART-JEPA, Global HEART-JEPA, same-view JEPA, and cross-view contrastive learning; each external model uses one. These checkpoint sets are reported separately from the intervention runs; their run-level correspondence is unspecified here. The available-view, label-efficiency, and external analyses use a separate single benchmark checkpoint, distinct from the matched three-run means.

Data-separation audit. The pretraining cohort comprises 39,992 unique participants. The 17,841 C2 study identifiers have zero overlap with the pretraining manifest at the visit and instance level. C2 evaluates held-out studies, and C3 evaluates participants absent from both pretraining subsets. Pretraining-participant exclusion, rather than IDP availability, defines C3.

Table S2: Matched target-mixture intervention. Values are mean ± sample SD over three training seeds, on C2. Retrieval is top-1 percentage over six ordered view directions with the full gallery; transfer is cross-view structural R^2 without target-view refitting; cardiac probing uses all 17 IDPs that test on C3.

p_{same}	Matched cosine	Transfer R^2	Retrieval top-1	Cardiac probe R^2
0	0.932 ± 0.021	0.479 ± 0.022	41.13 ± 5.49	0.743 ± 0.007
.25	0.254 ± 0.037	0.062 ± 0.104	0.207 ± 0.190	0.688 ± 0.016
.50	0.114 ± 0.025	-0.965 ± 0.524	0.0109 ± 0.0042	0.618 ± 0.054
.75	0.119 ± 0.007	-1.327 ± 0.235	0.0146 ± 0.0039	0.660 ± 0.011
1	0.140 ± 0.069	-1.039 ± 0.393	0.0277 ± 0.0101	0.677 ± 0.021

Table S3: Per-seed results for the endpoints of the matched target-source intervention. Retrieval is top-1 percentage; transfer and cardiac probing are R^2.

p_{same}	Seed	Matched cosine	Transfer	Retrieval	Cardiac probe
0	42	0.918	0.454	36.7683	0.734
0	43	0.956	0.485	39.3205	0.746
0	44	0.923	0.497	47.2937	0.748
1	42	0.081	−1.478	0.0234	0.664
1	43	0.215	−0.919	0.0392	0.701
1	44	0.124	−0.720	0.0206	0.665

D TARGET-QUERY AND GLOBAL HEART-JEPA RESULTS

Independent target queries preserve positive structural transfer and similar cardiac probing, with higher functional transfer in all three paired seeds (Tables S4 and S5). All results use frozen EMA encoders; probing and transfer use C2, and retrieval uses C3 queries against the fixed C2 gallery. The intervention removes exact context-mask index reuse while retaining incidental overlap (Jaccard 0.682 ± 0.072).

Global HEART-JEPA has stronger fitted cardiac probing and higher raw cosine similarity in all three seeds, while HEART-JEPA has a larger matched–mismatched margin and better participant retrieval, structural transfer, and functional transfer. Useful cardiac information therefore coexists with negative mean transfer in Global HEART-JEPA. Its stronger RVEF prediction in Table 2 further distinguishes fitting a new prediction model from transferring one unchanged. Head design and normalization vary jointly in this comparison, leaving their individual contributions unresolved.

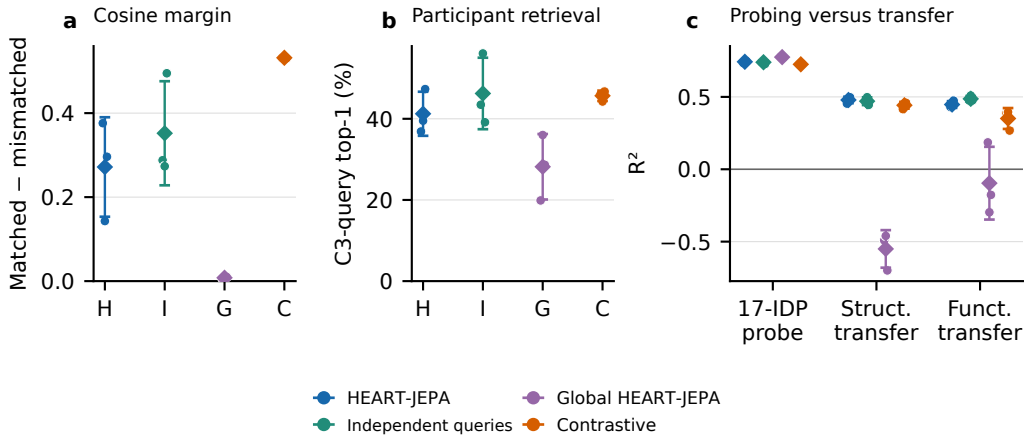


Figure S1: Cross-view objective controls. H: shared-query HEART-JEPA; I: independent target queries; G: Global HEART-JEPA; C: EMA-InfoNCE. Points are three training seeds and diamonds are mean ± sample SD. Margin and cardiac probes use C2; retrieval uses C3 queries with the fixed C2 gallery. The Global HEART-JEPA control has the higher 17-IDP probe score but negative mean structural and functional transfer. Head design and normalization differ between objective families.

Table S4: Target-relation controls (mean ± sample SD over seeds 42–44). Retrieval is top-1 percentage; cardiac probing and transfer are R^2.

Objective	Matched cosine	Retrieval (%)	Cardiac probe	Structural transfer	Functional transfer
HEART-JEPA	0.932 ± 0.021	41.24 ± 5.43	0.743 ± 0.007	0.479 ± 0.022	0.447 ± 0.023
Independent queries	0.923 ± 0.014	46.24 ± 8.81	0.739 ± 0.011	0.470 ± 0.027	0.487 ± 0.020
Global HEART-JEPA	0.988 ± 0.003	28.18 ± 8.07	0.775 ± 0.002	-0.550 ± 0.130	-0.096 ± 0.251

Table S5: Per-seed target-relation control results. Ret. and View ID are percentages; Probe, Struct., and Funct. are R^2. H denotes HEART-JEPA with shared query indices, I denotes independent target queries, and G denotes Global HEART-JEPA.

Control	Seed	Cosine	Margin	Ret.	View ID	Probe	Struct.	Funct.
H	42	0.918	0.3761	36.87	99.976	0.734	0.454	0.430
H	43	0.956	0.1430	39.52	99.979	0.746	0.485	0.438
H	44	0.923	0.2961	47.32	99.978	0.748	0.497	0.474
I	42	0.932	0.2878	43.47	99.974	0.742	0.469	0.475
I	43	0.930	0.2735	56.10	99.981	0.748	0.498	0.509
I	44	0.907	0.4949	39.15	99.979	0.728	0.445	0.475
G	42	0.987	0.0067	19.88	99.989	0.777	−0.492	0.186
G	43	0.991	0.0073	36.00	99.989	0.776	−0.460	−0.297
G	44	0.985	0.0099	28.65	99.991	0.772	−0.699	−0.177

E Matched Cross-View Contrastive Control

The three contrastive runs are paired by training seed with the pure cross-view HEART-JEPA runs. As above, retrieval uses C3 queries and the C2 gallery, while cardiac probing and direct transfer use C2. The comparison asks whether an objective optimized for participant discrimination also supports cardiac prediction across views.

Table S6: Matched objective-family comparison (mean ± sample SD over seeds 42–44). Retrieval is top-1 percentage; cardiac probing and transfer are R^2.

Objective	Margin	Retrieval	Cardiac probe	Structural	Functional
HEART-JEPA	0.272 ± 0.118	41.24 ± 5.43	0.743 ± 0.007	0.479 ± 0.022	0.447 ± 0.023
Contrastive	0.532 ± 0.001	45.67 ± 1.27	0.725 ± 0.002	0.442 ± 0.023	0.351 ± 0.072

HEART-JEPA has higher matched cosine in all three paired seeds (0.932 ± 0.021 versus 0.848 ± 0.007), while contrastive learning has the larger margin in all three. HEART-JEPA has higher cardiac-probe, structural-transfer, and functional-transfer scores in all three; contrastive retrieval is higher in two. View identity is nearly identical at $99.98 \pm 0.0019\%$ and $99.98 \pm 0.0039\%$.

F Participant-Disjoint Transfer and Representation Geometry

Table S7 reports the formal endpoints after excluding every evaluation participant present in the pretraining manifest. Centered geometry uses C4, identical across models; cardiac probes use C3, and retrieval uses C3 queries with the C2 gallery. Values are means and sample SD across seeds 42–44. Centering removes view-specific means for the geometric diagnostic; the transfer pipeline retains source-fitted standardization. Linear CKA characterizes full-feature geometry independently of a fitted target-view mapping. A linear source probe has predictions $w_c^\top D_c^{-1}(z^t - \mu_c) + b_c$, where D_c is its diagonal source scale matrix. A target feature-mean shift adds $w_c^\top D_c^{-1}(\mu_t - \mu_c)$ to the prediction mean and can reduce R^2 without changing Pearson correlation under a constant output shift. Transferred R^2 therefore combines calibration and predictive-information effects. Their relative contributions remain unresolved because the evaluation includes neither target-view recalibration nor transferred Pearson correlation.

Table S7: Participant-disjoint (C3) endpoint comparison. Retrieval is top-1 percentage with C3 queries and the fixed C2 gallery; transfer and cardiac probing are R^2. Linear CKA is invariant to orthogonal transforms and isotropic scaling.

Target	Centered cosine C4	Linear CKA C4	Retrieval C3 queries	Structural transfer C3	Cardiac probe C3
Cross-view	0.787 ± 0.027	0.850 ± 0.007	41.24 ± 5.43	0.477 ± 0.020	0.747 ± 0.007
Same-view	0.046 ± 0.021	0.055 ± 0.030	0.0279 ± 0.0103	-1.101 ± 0.393	0.680 ± 0.021

G Temporal Matching Across Sequence Positions

Ten temporal tubelets partition each fixed-length 50-frame input into normalized sequence positions. Inputs are truncated or repeat-last-padded, not temporally resampled. Sequence-position centering subtracts the across-participant mean token separately for each view and position. Table S8 averages each metric over six ordered view pairs within a seed and then reports mean $\pm$ sample SD over three seeds. The circular MAE is normalized by ten positions and treats sequence indices cyclically. It measures sequence-index correspondence rather than physiological phase error. Repeated terminal frames are a potential matching cue; padding prevalence and a phase-resampled control remain unmeasured. The relationship between this matching result and EF prediction has not been tested.

Table S8: Sequence-position-centered retrieval on C5. Top-1 chance is 10%.

Target	Top-1 (%)	Circular MAE	Same-index cosine	Index margin
Cross-view	68.49 ± 3.56	0.061 ± 0.011	0.757 ± 0.030	0.150 ± 0.034
Same-view	25.41 ± 2.55	0.196 ± 0.008	0.086 ± 0.009	0.057 ± 0.013

H Final-Checkpoint Objective Diagnostics

To characterize the sharp change near $p_{\text{same}} = .25$ discussed in Section 5.3, we evaluate each final checkpoint on fixed held-out batches under all same-view and cross-view relations, normalizing MSE by target-token variance. On a smaller fixed batch we also compute the cosine similarity between gradients induced by the same-view and cross-view losses. The pure cross-view endpoint has lower held-out cross-view error and more aligned same/cross local gradients. The measurements characterize local loss geometry at the final solution. Near-orthogonal gradients indicate weak local alignment, while the cause of the mixed-target performance drop requires evidence about the training trajectory.

Table S9: Final-checkpoint loss and gradient diagnostics (mean $\pm$ sample SD over three seeds). Gradient cosine is reported for the final encoder block and the predictor.

p_{same}	Cross-view NMSE	Same-view NMSE	Encoder grad. cosine	Predictor grad. cosine
0	0.298 ± 0.025	0.270 ± 0.014	0.469 ± 0.078	0.350 ± 0.048
.25	0.677 ± 0.014	0.371 ± 0.023	0.026 ± 0.005	0.042 ± 0.014
1	1.187 ± 0.039	0.552 ± 0.025	0.073 ± 0.024	0.068 ± 0.037

I Cardiac Phenotype Prediction and Transfer

Phenotypes and feature construction. The 17 IDPs follow the field definitions in the UK Biobank Data Showcase. The 13 structural measures comprise LV/RV end-diastolic and end-systolic volumes, LA/RA maximum and minimum volumes, LV myocardial mass, and four wall-thickness measures. The four functional measures are average heart rate, LA stroke volume, LA ejection fraction, and global LV circumferential strain; grouping stroke volume as functional is a reporting convention. All 17 fields receive equal weight in the overall cardiac-probe score, and the source-data catalogue supplies each field ID, unit, definition, and official URL. Continuous-IDP probes concatenate the

three frozen view representations. Disease classifiers use elementwise averaging, as specified in Appendix L.

I.1 Repeated-CV protocol for ventricular ejection fraction

LVEF and RVEF are evaluated separately from the 17-IDP mean and its 13-structural/four-functional groups. Table 2 uses C3 ($N = 15{,}935$) for all seven models. For each model, the frozen 2Ch, 3Ch, and 4Ch features are concatenated and a supervised standardized linear probe is fitted on the training participants of each outer fold. The three complete five-fold partitions are shared across models. Preprocessing fitted from features and the probe itself use only the corresponding training fold.

For model m, let K_m be its reported checkpoint count and let $\widehat{\boldsymbol{y}}_{m,j}^{(r)}$ collect the OOF predictions from all five test folds in repetition $r \in \{1, 2, 3\}$ for checkpoint j. Each participant contributes one held-out prediction per repetition and checkpoint. For either EF endpoint and each metric $T \in \{\mathrm{MAE}, \mathrm{PearsonR}\}$, the reported aggregation is

$$t_{m,j}^{(r)} = T\left(\boldsymbol{y}, \widehat{\boldsymbol{y}}_{m,j}^{(r)}\right), \qquad a_m^{(r)} = \frac{1}{K_m} \sum_{j=1}^{K_m} t_{m,j}^{(r)}, \tag{8}$$

$$\overline{a}_m = \frac{1}{3} \sum_{r=1}^{3} a_m^{(r)}, \qquad s_m = \sqrt{\frac{1}{3-1} \sum_{r=1}^{3} \left(a_m^{(r)} - \overline{a}_m\right)^2}. \tag{9}$$

Each metric is computed from pooled OOF predictions within a repetition, then averaged across checkpoint *metrics* while keeping their predictions separate. Table 2 reports $\overline{a}_m \pm s_m$, with MAE in EF percentage points and R the Pearson correlation. The sample SD measures sensitivity to the three partitions of the fixed cohort, conditional on the available checkpoints. It quantifies split variability rather than pretraining-seed variability.

Table 2 uses three checkpoints for each of the four internal models and one for each of the three external systems. The LVEF comparison between HEART-JEPA and Global HEART-JEPA is descriptive. All models share the evaluation cohort and folds; external pretraining and input adapters differ. C3 excludes participants in our pretraining manifest, while overlap with external pretraining manifests remains unverified.

Which phenotypes transfer across views? Figure S2 decomposes the matched endpoint comparison on C3. LV global wall thickness and myocardial mass have transfer scores of 0.609 ± 0.022 and 0.596 ± 0.021, followed by LA minimum volume at 0.569 ± 0.042, whereas RA maximum volume reaches 0.112 ± 0.042. For RA maximum and minimum volumes, 4Ch-to-3Ch transfer is negative (-0.284 ± 0.192 and -0.366 ± 0.267) despite positive direction-averaged scores. In standard long-axis planning, 3Ch follows the left ventricular outflow tract rather than the right atrium, whereas 4Ch includes both atria (Kramer et al., 2020). The 4Ch-to-3Ch RA test therefore changes both anatomical visibility and feature coordinates. These factors jointly frame the transfer result; the analysis does not isolate visibility as the cause, and indirect anatomical correlates can remain predictive. These exploratory analyses use the common crop prior without demographic or body-size adjustment.

Post-hoc left- and right-heart structural groups. To separate traits with different anatomical coverage, we re-aggregate the saved C3 per-trait scores into nine LV/LA traits (LV volumes, LV mass, four wall-thickness measures, and LA volumes) and four RV/RA volume traits. The groups classify phenotypes by chamber rather than by visibility in an individual slice. Within each of the three matched pretraining seeds, we average the already fold- and direction-averaged transfer R^2 equally across traits; Table S10 then reports the mean and sample SD across seeds. This summary reuses the per-trait results without refitting and is reported alongside the original 13-IDP structural mean.

Table S10: Post-hoc anatomical grouping of direct structural transfer on C3. Values are mean ± sample SD across three pretraining seeds after averaging traits within each group. All six ordered view directions are retained. These are descriptive summaries of traits with different anatomical coverage.

Target	LV/LA structural traits ($n = 9$)	RV/RA volume traits ($n = 4$)
HEART-JEPA	0.546 ± 0.021	0.322 ± 0.020
Same-view JEPA	-1.134 ± 0.395	-1.029 ± 0.400

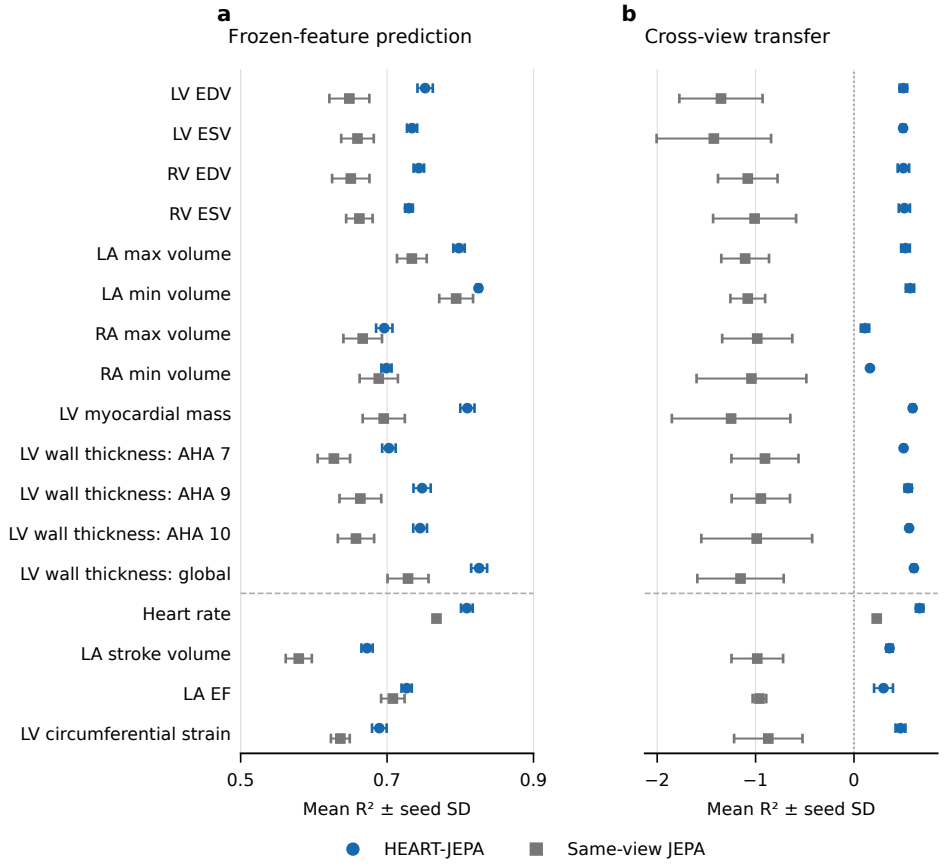


Figure S2: Cardiac phenotype prediction and transfer across anatomical views. All 17 verified fields are shown in anatomical order, with a dashed separator between structural and functional reporting groups. Points and bars show mean and sample SD across seeds 42–44 on C3. (a) A standardized OLS probe on concatenated three-view embeddings; five participant folds averaged within each seed. (b) A source-view standardized ridge probe ($\alpha = 1$) applied to a different view without refitting; five folds and six ordered directions averaged within each seed. Negative R^2 indicates worse prediction than the target test-set mean. Training seeds are the replication unit; traits, folds, and directions are correlated. Source data provide fold-level and directional values.

Table S11: Five-fold mean R^2 for cardiac IDPs and the across-17-IDP mean. *The Mean column averages all 17 IDPs: the nine shown here and the eight in Table S12.* All four displayed models use C2; RV/LV: right/left ventricle; RA/LA: right/left atrium; EDV/ESV: end-diastolic/end-systolic volume. Numbers below headers are UK Biobank field IDs. External models use their respective pretraining and input adapters (Section 5.6).

Model	LVEDV 24100	LVESV 24101	RVEDV 24106	RVESV 24107	LAmax 24110	LAmin 24111	RAmax 24114	RAmin 24115	LVmass 24105	Mean
HEART-JEPA	0.767	0.746	0.755	0.743	0.790	0.811	0.699	0.701	0.830	0.754
CMR Transformer	0.792	0.771	0.761	0.717	0.734	0.737	0.606	0.613	0.812	0.704
CineMA	0.805	0.716	0.780	0.739	0.833	0.791	0.786	0.748	0.840	0.689
ViTa-MAE	0.565	0.515	0.568	0.546	0.576	0.520	0.413	0.404	0.607	0.499

Table S12: Five-fold mean R^2 for the remaining cardiac IDPs, with the same cohort as Table S11 (four displayed models on C2). WT7/9/10/G: LV mean wall thickness in AHA segments 7, 9, 10, and globally; HR: average heart rate; LASV/LAEF: LA stroke volume/ejection fraction; GCS: global LV circumferential strain. Numbers below headers are field IDs.

Model	WT7 24130	WT9 24132	WT10 24133	WTG 24140	HR 22426	LASV 24112	LAEF 24113	GCS 24157
HEART-JEPA	0.725	0.774	0.762	0.846	0.799	0.668	0.706	0.690
CMR Transformer	0.618	0.683	0.689	0.752	0.785	0.645	0.598	0.650
CineMA	0.635	0.685	0.687	0.765	0.389	0.602	0.525	0.381
ViTa-MAE	0.433	0.471	0.492	0.550	0.603	0.439	0.347	0.433

J Directional Retrieval and Available-View Results

The separate single benchmark checkpoint described in Section 5.6 achieves 47.70%/68.77% top-1/top-5 retrieval over six ordered view directions (random top-1: 0.0056%). All scans are pretraining-disjoint; excluding these participants from both queries and gallery gives 48.74%/69.67%, and excluding them only from queries gives 47.69%/68.69%. For comparison, the matched three-seed cross-view retrieval mean is $41.13 \pm 5.49\%$.

Table S13: Same-participant retrieval on C2, averaged over six ordered view directions. External models are evaluated as complete pretrained systems with their respective input adapters.

Model	Top-1	Top-5
HEART-JEPA	47.70	68.77
CMR Transformer	0.50	1.74
CineMA	0.01	0.06
ViTa-MAE	0.02	0.07

Table S14: HEART-JEPA retrieval by ordered direction on C2.

Direction	Top-1 (%)	Top-5 (%)
2Ch→3Ch	43.16	65.29
3Ch→2Ch	45.87	67.46
2Ch→4Ch	39.85	62.14
4Ch→2Ch	41.77	63.60
3Ch→4Ch	58.72	77.45
4Ch→3Ch	56.84	76.65

Table S15: Available-view mean IDP R^2 on C2. A separate probe is fitted for each available-view subset using the same five participant folds. Each score measures prediction with a probe fitted for that input subset.

Available views	HEART-JEPA	CMR Transformer	CineMA	ViTa-MAE
2Ch+3Ch+4Ch	0.754	0.704	0.689	0.499
3Ch+4Ch	0.720	0.670	0.661	0.473
2Ch+4Ch	0.742	0.681	0.675	0.491
2Ch+3Ch	0.702	0.662	0.612	0.451
2Ch only	0.655	0.584	0.535	0.389
3Ch only	0.637	0.603	0.549	0.377
4Ch only	0.684	0.611	0.624	0.422

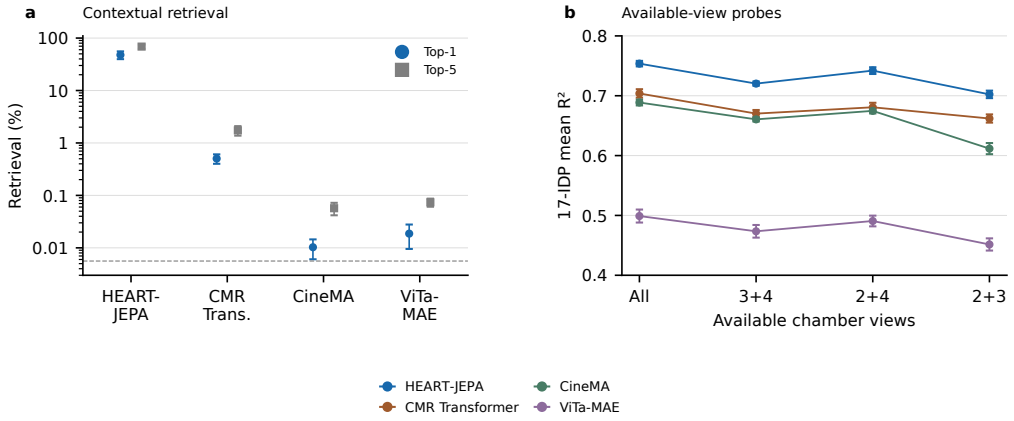


Figure S3: Contextual retrieval and available-view cardiac probing on C2. (a) Means and sample SD across six ordered view directions; the dashed line is uniform-random top-1. (b) Mean ± sample SD across five probe folds, after averaging 17 IDPs within each fold. Each input subset has a separately fitted probe. The benchmarks compare the supplied frozen representations under their respective input adapters.

K External Long-Axis Classification

M&Ms-2 evaluation uses C6 and frozen 4-chamber long-axis features. Seven disease-vs-normal classifiers each contrast one disease category against the 75 normal participants; eight one-vs-rest classifiers include the normal category and use all 360 participants. Both protocols use five stratified folds with split seed 42. A StandardScaler, up-to-64-component PCA, and class-balanced linear SVM ($C = 1$) are fitted inside each training fold, with no target-dataset encoder fine-tuning. Macro AUC averages class-specific fold-mean AUCs equally. We report the two protocols separately because their negative classes differ. These classification benchmarks are derived from the M&Ms-2 resource, whose official task is segmentation. Each model retains its own pretraining (Section 5.6).

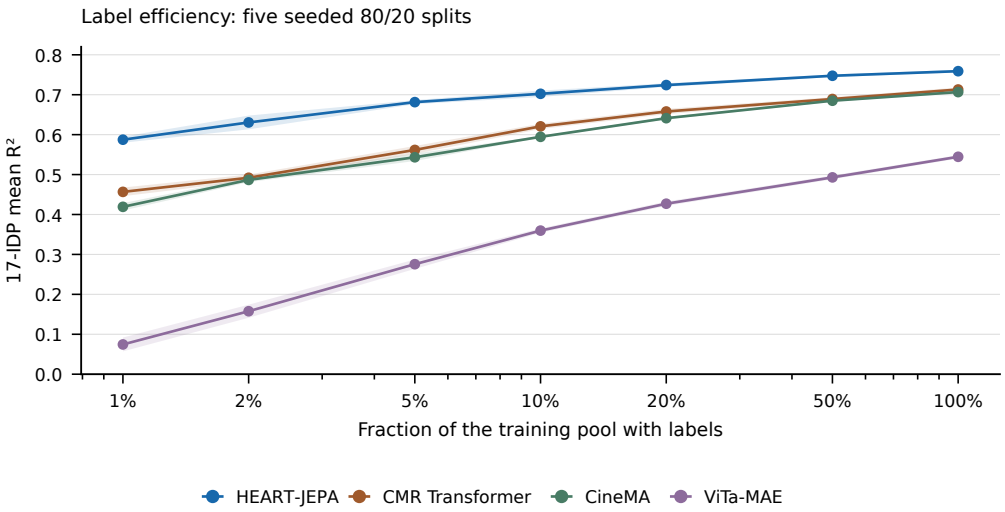


Figure S4: Frozen-feature label efficiency for 17 cardiac IDPs. Curves average five seeded 80/20 participant splits (seeds 0–4); bands show the saved population SD across splits (ddof $= 0$), measuring split variability for the evaluated checkpoints. Fractions refer to the 80% training pool; standardization and RidgeCV fitting use only that fraction, and the held-out 20% is fixed within each seed. This differs from the five-fold ordinary-linear-regression benchmark; its 100% endpoint summarizes the RidgeCV split protocol.

Table S16: Per-category external classification AUC on M&Ms-2 (C6) using frozen 4-chamber features. Values are mean ± sample SD across five stratified folds (split seed 42). n_+/n_- counts positive/negative participants before folding. Bold marks the highest displayed mean per row, including ties. Macro rows report averages of class means without pooled error bars.

(a) Disease vs normal Category	n_+/n_-	HEART-JEPA	CMR Transformer	CineMA	ViTa-MAE
ARR	35/75	0.703 ± 0.093	$\mathbf{0.758} \pm 0.087$	$\mathbf{0.758} \pm 0.096$	0.693 ± 0.142
CIA	35/75	0.878 ± 0.066	$\mathbf{0.901} \pm 0.014$	0.800 ± 0.138	0.800 ± 0.068
FALL	35/75	0.992 ± 0.012	$\mathbf{1.000} \pm 0.000$	0.989 ± 0.012	0.846 ± 0.037
HCM	60/75	$\mathbf{0.910} \pm 0.037$	0.894 ± 0.056	0.872 ± 0.037	0.808 ± 0.069
LV	60/75	0.852 ± 0.031	$\mathbf{0.950} \pm 0.065$	0.941 ± 0.023	0.808 ± 0.068
RV	30/75	0.696 ± 0.142	$\mathbf{0.818} \pm 0.148$	0.713 ± 0.063	0.629 ± 0.072
TRI	30/75	0.907 ± 0.066	$\mathbf{0.967} \pm 0.063$	0.878 ± 0.053	0.798 ± 0.094
Macro mean	–	0.848	**0.898**	0.850	0.769
(b) One vs rest Category	n_+/n_-	HEART-JEPA	CMR Transformer	CineMA	ViTa-MAE
ARR	35/325	0.615 ± 0.120	0.609 ± 0.166	0.593 ± 0.145	$\mathbf{0.644} \pm 0.094$
CIA	35/325	0.722 ± 0.080	$\mathbf{0.766} \pm 0.109$	0.714 ± 0.095	0.656 ± 0.146
FALL	35/325	$\mathbf{0.948} \pm 0.043$	0.931 ± 0.035	0.935 ± 0.033	0.840 ± 0.104
HCM	60/300	0.899 ± 0.050	$\mathbf{0.906} \pm 0.018$	0.863 ± 0.069	0.824 ± 0.076
LV	60/300	0.728 ± 0.056	$\mathbf{0.863} \pm 0.074$	0.836 ± 0.046	0.624 ± 0.085
NOR	75/285	0.752 ± 0.047	$\mathbf{0.900} \pm 0.030$	0.777 ± 0.046	0.658 ± 0.087
RV	30/330	0.605 ± 0.104	0.545 ± 0.091	$\mathbf{0.658} \pm 0.127$	0.601 ± 0.088
TRI	30/330	0.859 ± 0.011	$\mathbf{0.885} \pm 0.035$	0.783 ± 0.110	0.526 ± 0.190
Macro mean	–	0.766	**0.801**	0.770	0.672

Dataset category codes: ARR, arrhythmogenic cardiomyopathy; CIA, inter-atrial communication; FALL, tetralogy of Fallot; HCM, hypertrophic cardiomyopathy; LV/RV, dilated left/right ventricle; NOR, normal; TRI, tricuspid regurgitation. Fold SD is descriptive: training sets overlap and categories reuse participants.

L UKB DISEASE-LABEL PROBING

The HEART-JEPA, same-view, and cross-view contrastive checkpoints retain the shared backbone, pretraining data, and optimization budget; head designs differ across objective families. All three models use existing frozen features on C3; this benchmark is distinct from the EF protocol in Appendix I.1.

Participants and endpoints. The candidate cohort is C3. We use the earliest available scan with all three long-axis views for each participant. ICD-10-derived participant lists define five endpoints: atrial fibrillation (590 cases), atrioventricular block (195), diabetes (650), heart failure (189), and myocardial infarction (321). Each endpoint is evaluated against the same 14,362 participants absent from all seven disease lists. This control definition is based on recorded labels rather than a clinical assessment of health. Dilated and hypertrophic cardiomyopathy are excluded from the scored endpoints because of limited case counts but remain exclusion criteria for controls. Participants can contribute to more than one endpoint.

Frozen probes. For each scan we average the 384-dimensional 2Ch, 3Ch, and 4Ch features elementwise and use the result without PCA or truncation. Within each training fold we fit a StandardScaler followed by L2-regularized logistic regression ($C = 1$, class-balanced weights, LBFGS, at most 2,000 iterations). All models use identical participant-level five-fold stratified splits for split seeds 42, 43, and 44. Preprocessing and classifier fitting are restricted to each training fold. HEART-JEPA, same-view JEPA, and cross-view contrastive learning each have three pretrained checkpoints, evaluated on every split. Elementwise averaging gives a fixed 384-dimensional input rather than the 1,152-dimensional concatenation used for continuous IDPs and EF. It imposes symmetric mixing across views and can favor channel-compatible representations. The disease results are specific to this aggregation protocol; mean-versus-concatenation sensitivity remains untested.

Results. Table S17 reports ROC-AUC and average precision. The latter uses the average precision score rather than trapezoidal integration of the precision–recall curve. Values average folds, then training checkpoints, then split seeds, and the reported SD is across the three split-level means. HEART-JEPA has higher mean ROC-AUC and AP than same-view JEPA across all five disease endpoints. Against cross-view contrastive learning it has higher ROC-AUC for atrial fibrillation and heart failure and higher AP on four endpoints, with contrastive learning ahead on the remaining ROC-AUC endpoints and on AP for atrial fibrillation. Both cross-view objectives retain information relevant to the disease labels provided by UK Biobank.

These results quantify discrimination of the available UK Biobank disease labels. Diagnosis timing relative to MRI and the complete code-selection rules are absent from the label records, placing prospective risk prediction and clinical validation outside the scope of this evaluation.

Table S17: UKB disease-label probing with full 384-dimensional three-view mean features. Values are mean ± SD across three split-level means after averaging folds and training checkpoints. Same-view denotes the single shared-encoder $p_{\text{same}} = 1$ control used across all three views. AP is average precision. The models share endpoint-specific participants and folds.

Disease (n_+)	HEART-JEPA	Same-view	Contrastive
ROC-AUC			
Atrial fibrillation (590)	0.777 ± 0.005	0.752 ± 0.002	0.776 ± 0.004
AV block (195)	0.797 ± 0.004	0.753 ± 0.002	0.807 ± 0.006
Diabetes (650)	0.773 ± 0.003	0.758 ± 0.002	0.776 ± 0.003
Heart failure (189)	0.801 ± 0.007	0.752 ± 0.007	0.786 ± 0.010
Myocardial infarction (321)	0.754 ± 0.003	0.713 ± 0.002	0.756 ± 0.005
Average precision (AP)			
Atrial fibrillation (590)	0.291 ± 0.006	0.267 ± 0.001	0.301 ± 0.002
AV block (195)	0.131 ± 0.003	0.071 ± 0.002	0.103 ± 0.001
Diabetes (650)	0.176 ± 0.002	0.151 ± 0.005	0.172 ± 0.005
Heart failure (189)	0.162 ± 0.004	0.121 ± 0.007	0.136 ± 0.002
Myocardial infarction (321)	0.101 ± 0.004	0.080 ± 0.002	0.090 ± 0.006

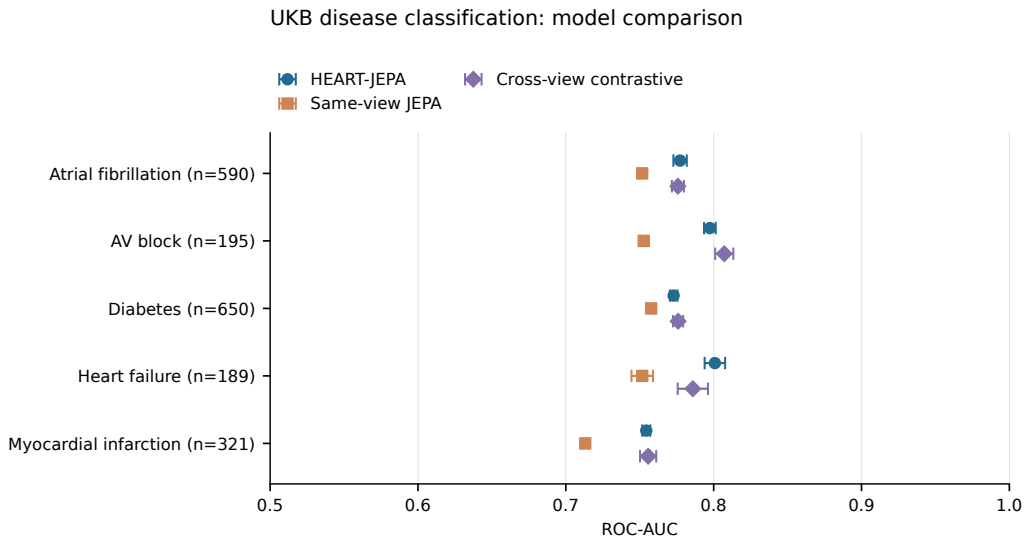


Figure S5: UKB disease-label discrimination from frozen three-view representations on C3. Each endpoint uses the same participants and test folds across models; n denotes cases, with 14,362 controls per endpoint. Points average five folds, three training checkpoints, and three split seeds. Error bars show SD across split-level means.