4D JEPA (Joint Embedding Predictive Architectures) for Cardiac MRI

Marta Hasny[1] Sameer Ambekar[1]

Prof. Dr. Julia Schnabel[1]

[1]*Chair for Computational Imaging and AI in Medicine (CompAI), TU Munich, Germany*

1 Introduction

Joint Embedding Predictive Architectures (JEPA) [1, 2, 5] offer a compelling alternative to pixel-reconstruction pretraining. Standard Transformer-based pretraining objectives, such as autoregressive next-token prediction in language models and pixel-level reconstruction in vision, are designed to recover the input signal. JEPA instead predicts the representation of a masked region in latent space, rendering the model invariant to high-frequency detail that carries no semantic content. Its video extension, V-JEPA [2], learns motion-sensitive representations by predicting across time, generalising to spatio-temporal modality.

Cine Cardiac Magnetic Resonance (CMR) [3] imaging is intrinsically four-dimensional: a three-dimensional volume of the heart captured across the cardiac cycle. Yet the field has not converged on how to present this structure to deep learning models. Simpler formats, such as 2D slices or key frames, are economical but discard motion or spatial context; richer formats, such as full 3D or 4D volumes, preserve this information at considerable computational cost. JEPA is well suited to this trade-off: rather than reconstructing pixel-level detail, it predicts how a representation evolves across space and time, aligning with clinical practice, where diagnosis depends on tracking wall motion and chamber volume rather than intensities corrupted by scanner noise. Yet no prior work applies this paradigm to cine CMR in its native four-dimensional form; existing approaches instead rely on key frames alone or embed the full sequence into architectures designed for 3D input, leaving open how much clinically relevant information is discarded. Reconstruction-based pretraining [4, 6] compounds this problem, as decoding full 4D volumes is where its computational cost is greatest. By predicting in latent space, JEPA circumvents this cost, and this thesis develops the first pretraining scheme to operate natively on full 4D cardiac data. What remains missing, then, is not evidence that pretraining benefits cardiac imaging, by now well established, but a principled account of how the data's own structure should even be used in the model.

2 Objectives

- The first comprehensive comparison of full 4D input against end-diastolic and end-systolic (ED/ES) key frames, architecture and data held fixed, isolating dimensionality as the sole determinant of any performance gap
- A standardized evaluation of 2D, 2D+T, 3D, and 4D CMR formats on segmentation and functional tasks under matched architectures, establishing which format justifies its computational cost
- Adapting Joint Embedding Predictive Architectures (JEPA) to 4D cine CMR to learn latent features without pixel-level reconstruction so that the model capacity targets cardiac motion rather than scanner noise.

3 Application

Kindly submit your application at the TUM Thesis Management Portal.

4 Requirements

- Good understanding of deep learning and, ideally, vision-language models.

- Practical experience with Python and PyTorch.
- Interest in medical imaging research; familiarity with representation learning is an advantage.

References

[1] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

[2] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024.

[3] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 2020.

[4] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[5] Y. LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[6] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022.