Joint Embedding Predictive Architecture (JEPA) for Cardiac Vision-Language Understanding

Marta Hasny[1] Sameer Ambekar[1] Cristina González[1]
Prof. Dr. Julia Schnabel[1]
[1]*Chair for Computational Imaging and AI in Medicine (CompAI), TU Munich, Germany*

1 Introduction

Transformer models trained to reconstruct their input, whether through next-token prediction in language or pixel reconstruction in vision [4, 6], treat every detail of the signal as equally important. In echocardiography, this assumption is particularly costly, since much of the signal consists of speckle and acquisition artifacts that are stochastic and diagnostically irrelevant. Reconstruction objectives therefore expend capacity on unpredictable noise, while token-level objectives penalize a clinically equivalent paraphrase as severely as a factual error, yet penalize an incorrect ejection fraction value by only a single token. Both thus optimize for surface fidelity, whereas cardiac assessment depends on the underlying physiology.

To mitigate this, Joint Embedding Predictive Architectures (JEPA) [1, 2, 5] predict the *representation* of masked content rather than the content itself, allowing the model to disregard unpredictable detail and concentrate on coherent structure. V-JEPA 2 [1] extends this principle to video, and VL-JEPA [3] further couples it with language by predicting text embeddings instead of tokens. In echocardiography, JEPA-based pretraining has recently achieved good performance across clinical tasks with strong robustness to acoustic perturbations, as latent prediction favors cardiac structure and motion over speckle.

Despite this, JEPA models in medical imaging are mostly limited to producing representations, and while VL-JEPA establishes that language generation is feasible within this paradigm, it remains unclear what latent prediction learns that next-token prediction does not, and how a decoder should exploit it. The core JEPA mechanism, inferring unobserved content from observed context, may capture clinically relevant structure absent from token-generative models, much as cardiologists infer function across the cardiac cycle and across views from incomplete acquisitions. Whether this context learning can be made accessible to a decoder, which cardiac tasks stand to benefit most, and what it reveals about latent prediction as an alternative to next-token generation are worthy questions to explore for the thesis.

2 Objectives

- **Learned representation analysis.** Characterize what latent prediction retains or discards relative to reconstruction and autoregressive objectives.
- **Leveraging context learning.** Use or design mechanisms that leverage JEPA context learning beyond frozen feature extraction.
- **Evaluation across relevant tasks.** Assess where these mechanisms outperform next-token prediction, including on tasks such as regression and visual question answering.

3 Application

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 and interest in research.

References

[1] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[2] 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.

[3] D. Chen, M. Shukor, T. Moutakanni, W. Chung, L. Yu, T. Kasarla, A. Bolourchi, Y. LeCun, and P. Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. In *ICLR*, 2026.

[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] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *NeurIPS*, 2017.