TF-JEPA: PREDICTIVE ALIGNMENT OF TIME–FREQUENCY REPRESENTATIONS WITHOUT CONTRASTIVE PAIRS

Michael Chaykowsky
Rivian and Volkswagen Group Technologies
Palo Alto, CA, USA

ABSTRACT

Learning generalizable representations from multivariate time series is challenging due to complex temporal dynamics, distribution shifts, and the difficulty of effectively designing contrastive pairs. We introduce TF-JEPA, a noncontrastive self-supervised method that leverages predictive alignment to integrate representations from the time and frequency domains without relying on negative sampling. TF-JEPA utilizes dual online time and frequency encoders, each paired with its own momentum-updated target encoder, embedding both views into a stable and unified latent space. Experiments on sleep EEG, gesture recognition, mechanical fault detection, and EMG classification demonstrate that TF-JEPA matches or surpasses contrastive and time frequency consistency baselines.

Track: Research

1 INTRODUCTION

Learning effective representations from time-series is challenging due to complex dynamics and label scarcity, especially in medical settings where annotation is costly (Ismail Fawaz et al., 2018; Gupta et al., 2021; Harutyunyan et al., 2019). Transfer learning has emerged as a powerful paradigm in time-series modeling, enabling pre-trained representations to generalize across domains (Ye & Dai, 2021). Time-series signals also possess a natural time-frequency duality that many representation learning methods have yet to fully exploit. This duality is particularly critical in physiological signals such as EEG (Zhang & Yao, 2021), where both spectral and temporal features are diagnostically relevant. These factors motivate self-supervised learning approaches capable of leveraging abundant unlabeled data and facilitating transfer across tasks.

Contrastive learning, the dominant paradigm, aligns augmented views (positive pairs) while repelling different samples (negative pairs) (Chen et al., 2020; van den Oord et al., 2019). However, applying contrastive learning to time-series is particularly difficult because suitable augmentations and negative-pair selection are challenging to design (Zhang et al., 2022; Wickstrøm et al., 2022). These methods are sensitive to augmentation choice and require large batch sizes or memory banks (Chen et al., 2020).

Recent non-contrastive approaches, notably the Joint Embedding Predictive Architecture (JEPA) (LeCun, 2023), have shown that strong representations can be learned without explicit negative pairs. Predictive objectives of this kind have not yet been systematically explored for timeseries data, where the natural dual view of time and frequency gives a compelling test bed. While Time–Frequency Consistency (TF-C) (Zhang et al., 2022) successfully aligns domains, it relies on computationally expensive cross-sample negatives and is sensitive to augmentation choices.

In this work, we introduce TF-JEPA (Time-Frequency Joint Embedding Predictive Architecture), a non-contrastive self-supervised framework that aligns time and frequency representations through prediction rather than contrastive repulsion.

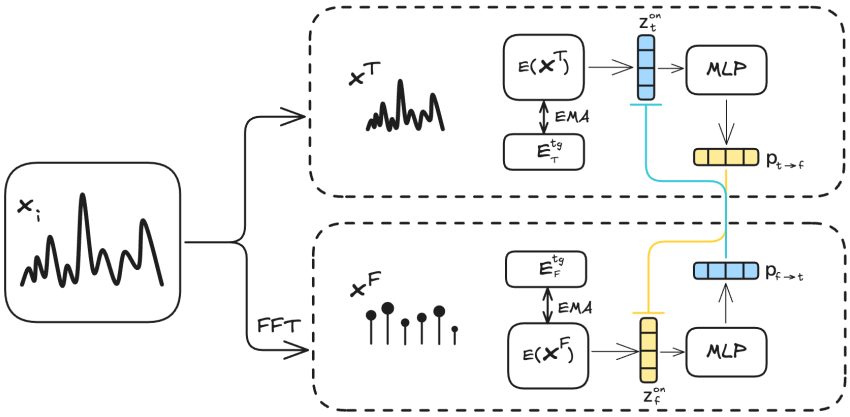

Figure 1: Architecture diagram for pre-training steps of TF-JEPA. This diagram communicates the three key ideas (i) time/frequency dual encoders, (ii) EMA targets, (iii) two cross-view predictors.

2 From TF-C to TF-JEPA

Time–frequency consistency (TF–C) established that aligning a waveform with its own spectrum can improve cross-dataset transfer in biosignal analysis. Yet TF–C depends on a contrastive objective whose computational and methodological demands have become increasingly restrictive. Contrastive learning requires large batches or memory queues, stores an $\mathcal{O}(B^2)$ similarity matrix, and, in practice, is vulnerable to "false negatives" in which two nearly identical signals are pushed apart.

Subsequent frequency-aware variants reduce some of these drawbacks but introduce bespoke components. Examples include masked frequency auto-encoders (Liu et al., 2024) and learnable Fourier filters, which rely on task-specific masking schemes that limit reuse.

TF–JEPA replaces the contrastive repulsion paradigm with predictive alignment, built on three design choices:

1. **Dual EMA targets.** A frozen time encoder and a frozen frequency encoder are updated after every step by an exponential moving average (EMA, momentum $m = 0.995$) of the online weights, providing stable target representations with no gradient overhead.
2. **Lightweight predictors.** Two small multilayer perceptrons, each mapping $\mathbb{R}^{128} \to \mathbb{R}^{128}$, transform the online embeddings so that they predict the corresponding target view. A BYOL-style (Grill et al., 2020) cosine loss aligns the two domains without negative pairs or large batch queues.
3. **End-to-end fine-tuning.** Because the objective avoids contrastive collapse, all encoder weights can be unfrozen during downstream training, allowing full adaptation to the target distribution (for example, SleepEEG $\to$ Epilepsy or HAR $\to$ Gesture).

TF–JEPA retains TF–C's intuition of cross-view alignment while reducing peak GPU memory by over $8\times$ during pre-training (Section 4.4) and improving cross-dataset transfer macro-F_1 by up to eight percentage points (for example, Fault Detection A $\to$B). Additional method intuition and analyses are provided in Appendix A.

3 Proposed Method

3.1 Model

Encoders. For every sample we form two views: a time-domain sequence $x_t \in \mathbb{R}^{B\times T\times C}$ and its frequency-domain counterpart $x_f = |\mathrm{FFT}(x_t)|$. Following TF-C[1], we compute a magnitude-only

[1]We note that while the TF-C paper describes using targeted single-component perturbations (E=1) with conditional boosting ($\alpha = 0.5$), their publicly available implementation uses a simpler approach that we adopt here for fair comparison.

Table 1: Transfer performance (%). **TS-TCC**[*], **TF-C**, and **TF-JEPA**[†] are pre-trained only on the single source dataset indicated (column 1) and then fine-tuned on the corresponding target dataset, following identical transfer-learning protocols. This setup allows direct comparison among three models of similar size. The right-most column reports the margin of TF-JEPA over the best competing transfer baseline on each task; positive values favor TF-JEPA.

Transfer task	TS-TCC				TF-C				TF-JEPA[†]				ΔF1	ΔAcc
	AUC	AP	Acc.	F1	AUC	AP	Acc.	F1	AUC	AP	Acc.	F1		
SleepEEG→Epilepsy	96.27	86.23	85.88	82.48	98.11	**94.56**	94.95	91.49	**99.07**	94.51	**95.31**	**92.24**	↑0.75	↑0.36
FD-A→FD-B	85.23	83.80	73.85	77.31	94.35	92.09	89.34	91.62	**99.98**	**99.47**	**99.28**	**99.47**	↑7.85	↑9.94
HAR→Gesture	86.60	65.61	63.33	59.91	89.55	65.91	68.33	65.79	**91.47**	**73.16**	**75.66**	**74.34**	↑8.55	↑7.33
ECG→EMG	**96.35**	**85.19**	85.88	**82.48**	87.53	82.74	85.37	80.51	92.53	79.41	**87.80**	80.03	↓2.45	↑1.92

spectrum over the full segment, with FFT size N equal to the sequence length defined in Appendix A. During pre-training, frequency augmentations randomly zero out or add noise to 10% of frequency bins, while time-domain augmentations apply jittering with $\sigma = 0.8$. Each view is processed by an identical L-layer one-dimensional Transformer encoder with model dimension d_{model}. After the Transformer, mean pooling over the temporal axis followed by a two-layer MLP projector produces latent vectors

$$z_t^{\text{on}},\ z_f^{\text{on}} \in \mathbb{R}^{d_z}, \qquad d_z = 128.$$

Momentum targets. Frozen target encoders G_t^{tg} (time) and G_f^{tg} (frequency) are updated after every optimization step by an exponential moving average (EMA) of the online encoder weights:

$$\theta^{\text{tg}} \leftarrow m\,\theta^{\text{tg}} + (1-m)\,\theta^{\text{on}}, \qquad 0.995 \le m \le 0.9995.$$

Because these target encoders are never back-propagated through, they add minimal memory and no optimizer state while outputting the reference embeddings z_t^{tg} and z_f^{tg}.

Predictors. Two lightweight predictor MLPs with dimensions $128 \to 256 \to 128$ are applied to the online embeddings. The time-view code is mapped to $p_{t\to f} = P_{t\to f}(z_t^{\text{on}})$ and trained to match the target frequency embedding z_f^{tg}. Symmetrically, the frequency-view code is mapped to $p_{f\to t} = P_{f\to t}(z_f^{\text{on}})$ and trained to match z_t^{tg}. Introducing such predictors, as in BYOL, helps stabilize training and prevents representational collapse.

3.2 Loss

The objective is the sum of two BYOL-style cosine regression terms,

$$\mathcal{L}_{\text{TF-JEPA}} = \mathcal{L}_{\cos}(p_{t\to f}, z_f^{\text{tg}}) + \mathcal{L}_{\cos}(p_{f\to t}, z_t^{\text{tg}})$$

where,

$$\mathcal{L}_{\cos}(p, z) = 2 - 2 \cdot \frac{p \cdot z}{\|p\|_2 \|z\|_2}$$

for each directional prediction. Maximizing cosine similarity aligns the two domains without requiring negative samples.

4 Experiments and Results

4.1 Experimental setup

We evaluate TF-JEPA on four widely-used cross-dataset transfer tasks in time-series representation learning. Each non-foundational model (TF-JEPA, TF-C, and TS-TCC (Eldele et al., 2021)) is pre-trained exclusively on the specified source dataset using the recommended hyperparameters from their respective papers, and then fine-tuned on the corresponding target dataset with identical classifier heads. All experiments were conducted on a single NVIDIA L4 GPU (24 GB memory) using mixed-precision training.

Table 2: Target-task performance (%) comparison against foundation models. NormWear is pre-trained on diverse wearable modalities. CBraMod is pre-trained on the full TUH EEG v2.0.1 corpus. Whereas, TF-JEPA is pre-trained on a 13.3% subset of TUH EEG v2.0.1.

Dataset	NormWear				CBraMod				TF-JEPA			
	AUC	AP	Acc.	F1	AUC	AP	Acc.	F1	AUC	AP	Acc.	F1
Epilepsy	**98.21**	**99.42**	95.51	92.61	98.02	99.10	90.35	97.23	97.53	98.90	**95.65**	**97.32**
FD-B	**84.54**	67.15	58.30	61.56	71.14	64.75	**75.49**	**65.58**	75.98	**71.16**	64.96	63.93
Gesture	88.56	64.33	55.00	49.04	**92.34**	**77.89**	**74.17**	**73.56**	89.75	62.83	62.50	57.32
EMG	93.73	83.85	87.71	62.39	**99.83**	**99.46**	**98.04**	**97.64**	95.51	92.28	84.67	84.88

4.2 Transfer learning performance

1. **SleepEEG→Epilepsy.** Transfer from 82 healthy overnight EEG recordings to seizure detection in 500 subjects.
2. **FD-A→FD-B.** Bearing-fault detection across two operating regimes with different torque and speed, testing robustness to mechanical covariate shift.
3. **HAR→Gesture.** Daily full-body motions (50 Hz, nine channels) to fine-grained hand gestures ($\approx$ 100 Hz, three channels), probing scale and granularity gaps.
4. **ECG→EMG.** Cross-organ physiological transfer: single-lead cardiac rhythms (300 Hz) to tibialis-anterior electromyograms (4 kHz).

TF–JEPA surpasses contrastive methods on SleepEEG→Epilepsy and on both domains of the Fault Detection benchmark and Gesture recognition, improving macro-F_1 by more than eight percentage points 1. TF-JEPA falls slightly short in the cross-organ physiological transfer task and a deeper analysis is shown in Appendix A.

4.3 Foundation model comparisons

For our foundation model benchmarks (Table 2), we provide reference for large-scale physiological priors. NormWear (Luo et al., 2024) is pre-trained on diverse wearable modalities; CBraMod (Wang et al., 2025) is pre-trained on a large EEG corpus using criss-cross spatial–temporal attention and masked patch reconstruction. TF-JEPA is pre-trained on a 13.3% subset of TUH EEG v2.0.1 (Neural Engineering Data Consortium (NEDC)) (groups 000–019) to keep comparisons similar. All models are fine-tuned only on each target dataset under identical heads and optimizers, without source→target transfer, serving as adaptation context for TF-JEPA. TF-JEPA outperforms both NormWear and CBraMod in accuracy and F1 metrics on the Epilepsy target task despite being pre-trained on significantly less data. This highlights the high data efficiency of our approach and presents an exciting avenue for future research into scaling these representations.

4.4 Resource usage

Because TF-JEPA replaces the contrastive NT-Xent objective with a non-contrastive BYOL-style cosine loss, it eliminates the quadratic $2B \times 2B$ similarity matrix that NT-Xent must materialize and back-propagate through at every pre-training step. Table 3 quantifies the effect on an NVIDIA L4 GPU with 178-step EEG windows and a batch size of 128: TF-JEPA allocates a peak of only 51 MB versus 421 MB for TF-C (an $8.2\times$ reduction) and completes each pre-training step in 26.7 ms versus 36.0 ms, a $1.35\times$ wall-clock speed-up.

A step-level breakdown reveals where the savings originate. Figure 2 shows the forward pass of TF-JEPA is marginally slower (12.7 ms vs. 11.0 ms) because it maintains a momentum-updated target encoder that doubles the number of stored parameters (2.49 M total vs. 1.18 M for TF-C, though only 1.31 M are trainable). The backward pass, however, is $1.8\times$ faster (14.1 ms vs. 25.0 ms), because the cosine loss produces a compact, $O(B)$ gradient graph in place of the $O(B^2)$ graph generated by the contrastive similarity matrix. This backward-pass saving more than compensates for the extra forward-pass cost, yielding a net reduction in both time and memory.

Table 3: Pre-training GPU footprint on an NVIDIA L4 (batch size = 128, sequence length = 178). All timings are medians over 105 profiled steps after a 5-step warm-up.

Metric	**TF-JEPA**	**TF-C**	**Ratio**
Total parameters	2.49 M	1.18 M	2.1×
Trainable parameters	1.31 M	1.18 M	1.1×
Peak GPU memory (MB)	51	421	0.12×
Avg. step time (ms)	26.7	36.0	0.74×
Forward (ms)	12.7	11.0	1.15×
Backward (ms)	14.1	25.0	0.56×

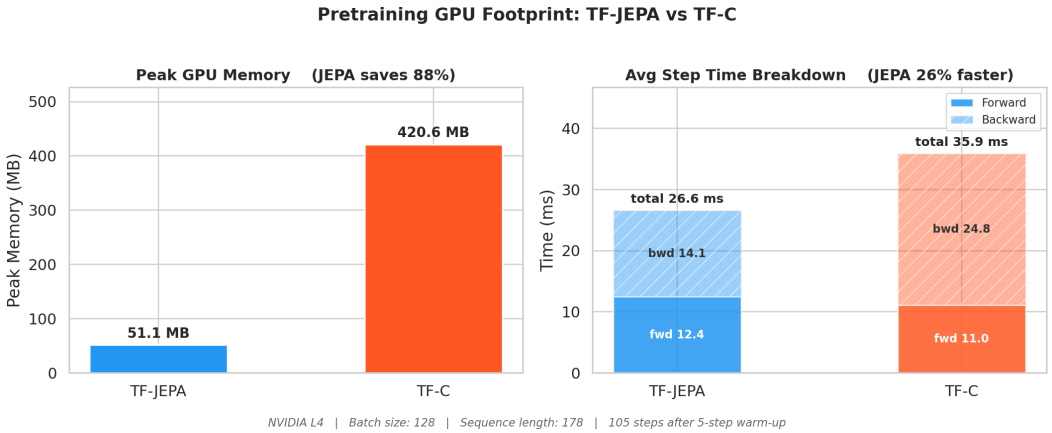

Figure 2: Pre-training GPU footprint comparison between TF-JEPA and TF-C on the SleepEEG pipeline (NVIDIA L4, batch size = 128). Peak memory and forward / backward time breakdown.

5 Conclusion

This work introduces TF-JEPA, a predictive, non-contrastive framework for learning shared time–frequency representations from unlabeled time-series data. By coupling an online time encoder with a momentum-updated frequency encoder and training them with a lightweight cosine loss, TF-JEPA removes the need for negative pairs and eliminates the quadratic cost of the contrastive similarity matrix. Controlled profiling on an NVIDIA L4 shows that this architectural choice reduces peak GPU memory by over $8\times$ relative to TF-C while delivering a $1.35\times$ wall-clock speed-up per pre-training step, despite TF-JEPA carrying roughly twice as many total parameters due to its momentum target network. On downstream transfer benchmarks, TF-JEPA improves cross-dataset performance by as much as eight percentage points. Because the objective is stable without a contrastive repulsion term, all encoder weights remain trainable during downstream fine-tuning, enabling full adaptation to target distributions.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc.,

2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn, 2017. URL https://arxiv.org/abs/1705.08168.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning, 2022. URL https://arxiv.org/abs/2204.07141.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. URL https://arxiv.org/abs/2105.04906.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020. URL https://arxiv.org/abs/2006.10029.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2352–2359, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Priyanka Gupta, Pankaj Malhotra, Jyoti Narwariya, Lovekesh Vig, and Gautam Shroff. Transfer learning for clinical time series analysis using deep neural networks, 2021. URL https://arxiv.org/abs/1904.00655.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time-series data. *Scientific Data*, 6(1):1–18, 2019.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Transfer learning for time series classification. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1367–1376, 2018. doi: 10.1109/BigData.2018.8621990.

Yann LeCun. A path towards autonomous machine intelligence. *Tech. Rep., Meta AI*, 2023. White paper.

Ran Liu, Ellen L. Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals, 2024. URL https://arxiv.org/abs/2309.05927.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals, 2024. URL https://arxiv.org/abs/2412.09758.

Neural Engineering Data Consortium (NEDC). Tuh eeg corpus (tueg) v2.0.1. Dataset release. URL https://isip.piconepress.com/projects/nedc/data/tuh_eeg/tuh_eeg/v2.0.1/. Accessed: 2026-01-27.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information, 2019. URL https://arxiv.org/abs/1905.06922.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs, 2021. URL https://arxiv.org/abs/2102.06810.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding, 2025. URL https://arxiv.org/abs/2412.07236.

Kristoffer Wickstrøm, Michael Kampffmeyer, Karl Øyvind Mikalsen, and Robert Jenssen. Mixing up contrastive learning: Self-supervised representation learning for time series. *Pattern Recognition Letters*, 155:54–61, March 2022. ISSN 0167-8655. doi: 10.1016/j.patrec.2022.02.007. URL http://dx.doi.org/10.1016/j.patrec.2022.02.007.

Rui Ye and Qun Dai. Implementing transfer learning across different datasets for time series forecasting. *Pattern Recognition*, 2021.

Xiang Zhang and Lina Yao. *Deep learning for EEG-based brain–computer interfaces: Representations, algorithms and applications*. World Scientific, 2021.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 3988–4003. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/194b8dac525581c346e30a2cebe9a369-Paper-Conference.pdf.

A Appendix

A.1 Why predictive alignment? Intuition behind TF–JEPA

Time and frequency as complementary "modalities". A discrete time–series and its Fourier spectrum form two loss-less, invertible views of the same signal. Similar to image–text pairs in CLIP (Radford et al., 2021) or audio–visual pairs in AVID (Arandjelović & Zisserman, 2017), these dual views emphasize different statistical regularities: the time domain exposes local temporal dynamics (e.g., waveform shape, transients), whereas the frequency domain highlights global rhythmic structure and stationarity. Leveraging both views therefore offers a built-in multi-modal supervision signal without requiring paired datasets from different sensors.

From contrastive repulsion to cross-view prediction. Contrastive objectives enforce invariance by repelling all other samples in the mini-batch, which costs $\mathcal{O}(B^2)$ memory and can mistreat near-duplicates as negatives. Joint-Embedding Predictive Architectures (JEPA) (LeCun, 2023) invert that idea: each online encoder predicts the latent vector produced by a slow-moving EMA target encoder of the opposite view. Concretely, the time encoder E_t^{on} learns to match the frequency target $z_f^{\text{tg}} = E_f^{\text{tg}}(x_f)$, while the frequency encoder E_f^{on} predicts the time target $z_t^{\text{tg}} = E_t^{\text{tg}}(x_t)$. This removes the need for negatives, keeps memory linear in B, and, like BYOL (Grill et al., 2020), prevents collapse because the EMA targets evolve slowly yet non-trivially. Applying JEPA across time/frequency views yields three benefits

1. **Semantic alignment.** Predicting one view from the other forces the network to focus on view-invariant factors (sleep stage, bearing damage, gesture identity) while disregarding nuisance details specific to either domain.
2. **Stability without collapse.** EMA targets provide a non-trivial prediction signal that evolves slowly; empirical and theoretical analyses (Tian et al., 2021; Bardes et al., 2022) show this circumvents trivial-solution collapse even with small batches.
3. **Linear complexity.** No $B \times B$ similarity matrix or memory queue is formed, so memory and compute scale linearly with B.

Why alignment should emerge self-supervised. Because the FFT is invertible, all task-relevant information in one view is present in the other. Minimizing the cosine distance between predicted and target embeddings therefore bounds the mutual information between the views from below (Poole et al., 2019); the optimum is reached when each encoder concentrates that shared information into its latent code. In practice we observe that the resulting representations cluster by semantics across datasets, echoing the theoretical expectation that view agreement acts as an information bottleneck selecting factors that generalize across domains.

Relation to prior multi-modal JEPA work. Concurrent studies have applied predictive objectives to RGB–depth pairs (Assran et al., 2022) and image–audio pairs (Alayrac et al., 2022). TF–JEPA is the first to exploit the intrinsic duality of a single signal, requiring no additional sensors or annotators. This property makes the method attractive for domains (e.g. medical telemetry, vibration monitoring) where extra modalities are costly or infeasible to collect.

A.2 Additional experimental details

This assertion is confirmed with an ablation study across six batch sizes from 16 to 512. For example, in the HAR transfer experiment TF-JEPA demonstrates robust performance across all batch sizes with a coefficient of variation of 2.05%, and accuracy saturating at around 76% for batch sizes ≥ 64. Notably, even at batch size 128 TF-JEPA allocates $8.2\times$ less peak memory than TF-C, while smaller batch sizes widen this gap further because the contrastive $O(B^2)$ cost shrinks faster than the fixed overhead of the momentum target network.

We further include CBraMod (Wang et al., 2025), a newly introduced brain foundation model for EEG decoding. Similar to NormWear, CBraMod is first pre-trained on a large heterogeneous corpus using a criss-cross Transformer backbone with parallel spatial-temporal attention and conditional masked EEG reconstruction on patch tokens. In particular, CBraMod is pre-trained on Version 2.0.1 of the TUH EEG dataset(Neural Engineering Data Consortium (NEDC)), which is organized

Table 4: Dataset statistics. C = number of classes after any relabelling; S = sampling rate; N_{pre} / N_{ft} give pre-training and fine-tuning sample counts. Window lengths follow cited preprocessing protocols.

Dataset	Domain	C	S (Hz)	Window	N_{pre}	N_{ft}
SleepEEG	EEG (sleep)	5	100	200	371 055	–
Epilepsy	EEG (seizure / normal)	2	178	178	–	60
FD-A	Vibro-acoustic (cond. A)	3	64 k	5 120	18 882	–
FD-B	Vibro-acoustic (cond. B)	3	64 k	5 120	–	18 864
HAR	9-axis IMU (daily activity)	6	50	128	10 299	–
Gesture	3-axis accel. (hand motion)	8	$\sim$100	256	–	440
ECG	Cardiac rhythm	4	300	1 500	8 528	–
EMG	Tibialis-anterior EMG	3	4 000	1 500	–	163

into roughly 150 patient groups (about 100 patients per group), providing broad clinical diversity and scale for learning generalizable EEG priors. Following the same downstream adaptation role as NormWear, CBraMod is then fine-tuned solely on each target dataset. This provides a second foundation reference point that measures how well generalized priors from large-scale EEG-only pre-training can adapt to downstream decoding tasks under the same target fine-tuning protocol used for NormWear, allowing TF-JEPA to be contextualized against both broad wearable pre-training and large EEG-only pre-training priors, while our TF-JEPA pre-training uses only TUH EEG v2.0.1 groups 000–019 to match the TUH data source at a smaller scale.

As shown in Figure 3, we notice that performance improves with higher EMA momentum m: we observe a positive correlation between m and transfer metrics (Pearson $r = 0.833$ across settings), with all metrics peaking at $m = 0.9995$. With 3 seeds for each m and a 95% CI on ΔF1, the best setting ($m = 0.9995$) exceeds the worst by $+11.3$pp in the HAR transfer experiment. This pattern generalizes across datasets: ECG shows the most dramatic sensitivity with a 39 percentage point improvement (53.7% $\rightarrow$ 92.7% accuracy), while SleepEEG exhibits optimal performance at the slightly lower $m = 0.995$ (90.8% accuracy). The dataset-dependent optimal momentum suggests that signal complexity influences the required target network stability. Biomedical time series with intricate temporal patterns (ECG, HAR) benefit most from ultra-slow updates ($m = 0.9995$), while sleep data achieves peak performance with moderate stability ($m = 0.995$). Intuitively, ultra-slow target updates stabilize the non-contrastive objective, improving stability and the signal-to-noise ratio in the target representations. The consistent superiority of high momentum values ($m \geq 0.995$) across all datasets validates the critical importance of target network stability in BYOL-style self-supervised learning for time series, with the EMA update rate of 0.05% or less proving optimal for complex temporal patterns.

A.3 Analysis of the ECG transfer case

The ECG$\rightarrow$EMG transfer has three classes labeled 0, 1, and 2. As shown in Figure 4, TF-JEPA identifies class 2 reliably but frequently predicts label 1 when the ground truth is 0, leading to the observed macro-F_1 drop. Classes 0 and 1 differ mainly by subtle waveform-shape variations; the explicit repulsion term in TF-C appears to preserve this fine boundary, whereas TF-JEPA's predictive loss focuses on cross-view alignment and is less sensitive to inter-sample separation. Introducing a class-balanced sampling during fine-tuning may help recover this distinction, and we leave that exploration to future work.

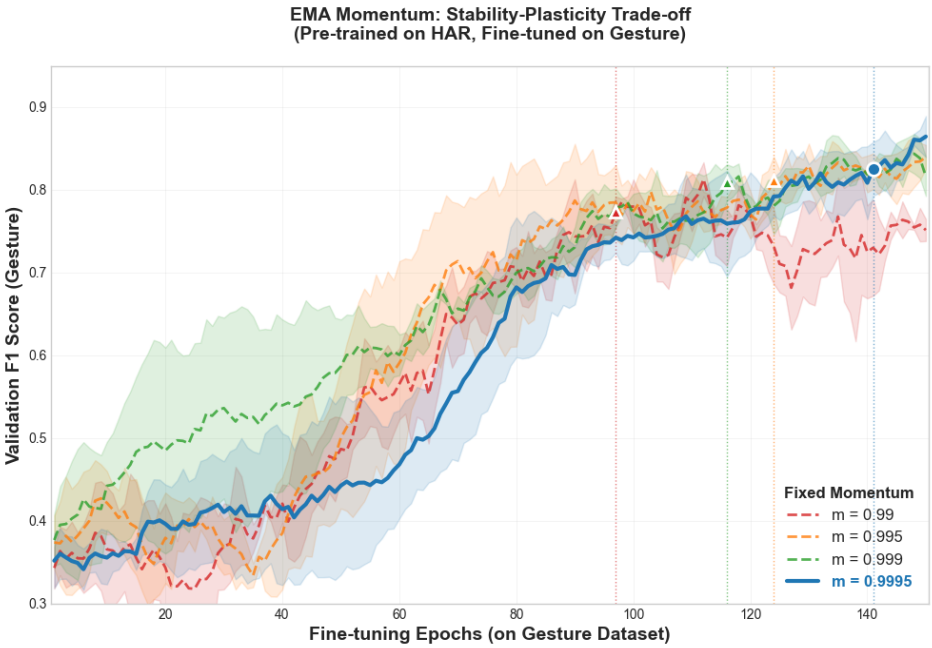

Figure 3: Validation F1 on Gesture (fine-tuning) after pre-training on HAR with fixed EMA momenta. Dotted lines show with 3 seeds for each m and 95% arrival epochs, $m = 0.9995$ converges more slowly than lower m but yields the highest final score, so we adopt it when final accuracy is prioritized over time-to-stability.

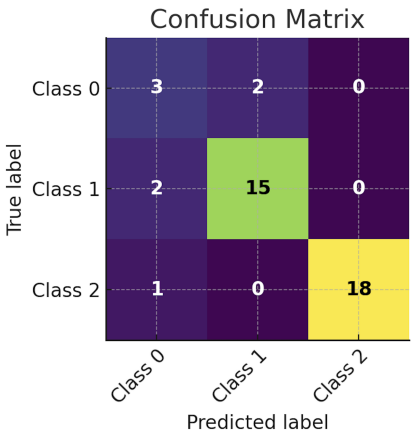

Figure 4: Confusion matrix for the 3-class test set (41 samples). Diagonal cells give correct predictions: class 0: 3/5, class 1: 15/17, class 2: 18/19 while off-diagonal counts expose the main failure mode. Class 0 & class 1 confusions (2 + 2 cases). Color intensity scales with sample count for quick visual emphasis.