Learning the Convolutional Basis for Time Series Clustering with a Temporal World Model

Sai Prasanna Teja Reddy Bogireddy (bogireddyteja@ChicagoBooth.edu)

CHICAGO BOOTH — The University of Chicago Booth School of Business

exelon℠

The Problem

Augmentation-built positives

Effective augmentations are domain dependent; they encode only self-similarity (a positive can never link two genuinely similar but distinct series), and they add hyperparameters that cannot be validated without labels.

Random kernels are never adapted

Random convolutional banks read out by PPV pooling are a strong, inexpensive basis, and R-Clustering clusters directly on them. Yet those kernels never see the data.

Our answer

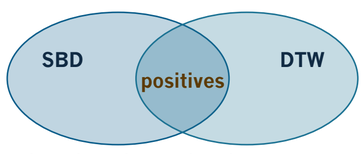
Consensus between complementary distance measures supplies the contrastive signal, and a temporal world model adapts the basis itself. No labels are required at any stage, and the PPV readout is unchanged.

Contributions

1. **A learned convolutional basis.** Trained without labels by causal latent prediction against a momentum target, keeping the cheap PPV readout.
2. **Consensus pairs, not augmentation.** Positive only if two complementary distances agree it is close; hard negative only if both agree it is far.
3. **One backbone, two roles.** World model and time encoder are one network, so both losses shape one representation.
4. **Fusion without a teacher metric.** A same-instance loss aligns the two views using neither labels nor distances.

Consensus Pair Mining

Each N × N distance matrix is precomputed once offline; training slices only the mini-batch sub-block. Two complementary measures per domain:



accepted only where both measures agree

$$P_i = N^{d_1}_{k_p}(i) \cap N^{d_2}_{k_p}(i)$$
$$N_i = \left(N^{d_1}_{k_n}(i) \cup N^{d_2}_{k_n}(i) \right)^c$$

The ambiguous remainder is discarded and excluded from the loss denominator, trading recall for high-precision positives.

SBD only	rigid phase shift	0.398
cDTW only	elastic warp, w = 0.1T	0.391
SBD ∪ cDTW	either measure agrees	0.383
LSD only	fine magnitude	0.401
CEP only	smoothed envelope	0.394
LSD ∪ CEP	either measure agrees	0.386
Consensus, both domains		**0.415**

Mean NMI when the mining rule is degraded in one domain at a time. Either measure alone trails consensus, and the union trails both: admitting a pair whenever either agrees inherits both teachers' blind spots.

A Learnable Basis

The first layer of the time encoder is a wide bank of M = 256 short kernels (length 9) over dilations {1, 2, 4, 8}, applied causally so no future leaks into a position. Kernels start zero-mean Gaussian and are trainable: world-model gradients flow into them.

The Temporal World Model

A causal backbone turns the bank responses into token states. The online encoder predicts, in latent space, the future token states of a momentum copy of itself across horizons Δ = 1 … H. The arrow of time is the only supervision, and nothing is reconstructed in input space.

$$L_{wm} = \Sigma_{i,\Delta} \left\| g\left(c^{\theta}_{1:L-\Delta}, e_{\Delta} \right) - \mathrm{sg}\left(c^{\xi}_{1+\Delta:L} \right) \right\|_{\text{smooth-}\ell 1}$$

Three ingredients keep this from collapsing to a constant:

the stop-gradient on the target,

the slow EMA, so the target never chases the online network,

distinct targets per horizon, forbidding a single trivial fixed point.

Readout and Clustering

The network is then frozen. The bank is applied non-causally for the readout (centered padding avoids a left-edge artifact), and each kernel gets a bias at a random quantile of its own response.

$$r_k(x) = (1 / T) \Sigma_u \; 1[\, (B_k * x)(u) > b_k \,]$$

The joint embedding is standardized, reduced by PCA at 95% of variance without whitening, then clustered by Euclidean k-means with K the known class count. Whitening would swamp the clusters.

The novelty is not the pooling, adopted unchanged — it is that the kernels being pooled were learned, not sampled at random.

Training

Bank	256 kernels, length 9, dilations {1, 2, 4, 8}
Backbone	stride 2, width D = 128, dilated causal residual blocks
World model	H = 4 horizons, smooth-$\ell 1$, EMA 0.996 → 1
Frequency enc.	1-D conv 64 / 128, embedding 128, GroupNorm
Contrastive	τ = 0.5, k_p = 5, k_n = 20, heads 2-layer MLP dim 64
Optimization	Adam, lr 10^{-3}, wd 10^{-4}, clip 5, 100 epochs, batch 256
Clustering	PCA 95% (no whitening), k-means, 10 inits
Runtime	7025 s per dataset, ≈ 1.7× the costliest baseline

Stable under one-at-a-time sweeps: the bank saturates by M = 256, horizons peak at H = 4, and k_p, k_n and τ each move NMI by at most 0.011. Every configuration still outranks the strongest baseline.

Results

All 128 UCR datasets, train and test merged as in prior work, over 5 seeds. Five baselines across three families.

0.415 mean NMI, best of all

54 / 128 outright wins, 2.6× the next best

2.43 average rank, best of all

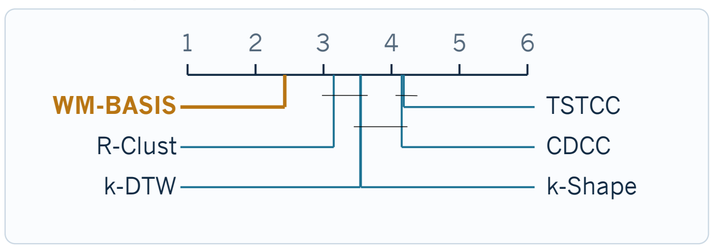
Method	NMI	RI	ARI	Rank	#Best	p (Holm)
k-DTW	0.333	0.723	0.246	3.54	12	4.0×10^{-9}
k-Shape	0.345	0.735	0.263	3.55	21	1.6×10^{-7}
TSTCC	0.292	0.684	0.200	4.18	11	3.9×10^{-13}
R-Clust	0.356	0.733	0.273	3.15	17	7.4×10^{-7}
CDCC	0.300	0.724	0.226	4.15	13	3.9×10^{-13}
WM-BASIS (ours)	**0.415**	**0.762**	**0.342**	**2.43**	**54**	**ref.**

Table 1 · Mean over all 128 UCR datasets; the last column is the Holm-corrected Wilcoxon p-value against that baseline.

Margins by family: +0.059 NMI over R-Clust, which shares our readout but keeps a random bank; +0.123 over TSTCC and +0.115 over CDCC; +0.082 and +0.070 over k-DTW and k-Shape.

Statistical Significance

Significantly better than every baseline on NMI at the 0.01 level, and the same on RI. A Friedman test rejects equal performance, $\chi^2 = 80.0$, $p = 8.5 \times 10^{-16}$.



Nemenyi diagram of average NMI ranks. WM-BASIS is first and joined to no baseline.

What Drives the Gain?

Two additions over a random-bank baseline are disabled independently: a world model that learns the bank, and the time–frequency fusion.

Variant	NMI	RI	ARI	Δ NMI
WM-BASIS (full)	**0.415**	**0.762**	**0.342**	—
w/o world model	0.366	0.744	0.283	−0.049
w/o fusion	0.385	0.749	0.310	−0.030
w/o both	0.354	0.741	0.278	−0.061

Removing both leaves a random bank read out by PPV, reproducing R-Clust (0.354 vs. 0.356) — a positive control. With fusion removed, learning the bank rather than fixing it lifts NMI to 0.385; together the two add +0.061, more than their separate sum.

Limitations

Distance precomputation is quadratic in N and dominates the offline cost, though the matrices are computed once and reused across every seed and sweep, and parallelize trivially. Very short series can leave no valid horizon in a batch; the term is then skipped. Univariate only.