ROBUST MULTIMODAL REPRESENTATION LEARNING THROUGH TOPOLOGY-PRESERVING PROTOTYPE MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

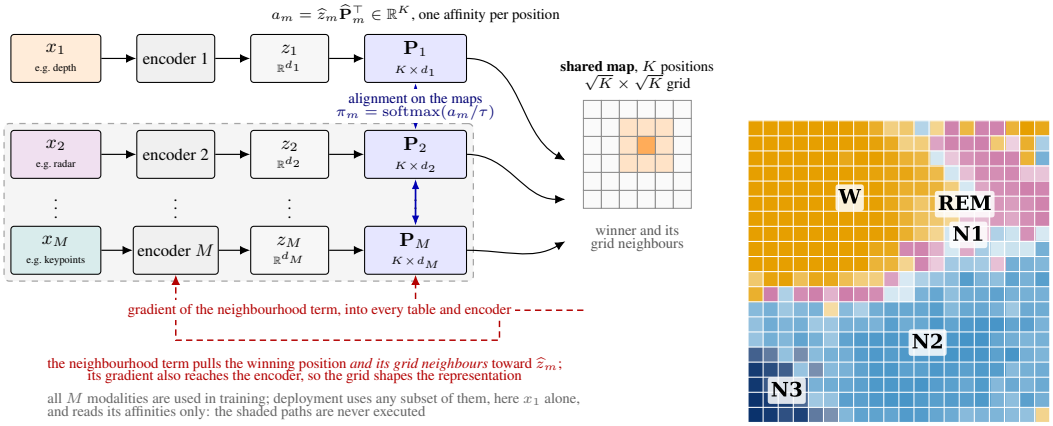
Learning representations that preserve relations between samples is a fundamental challenge in machine learning. For multimodal data, the representations should retain the information of each modality while remaining aligned across modalities. A common approach is to map all inputs into a shared latent space. However, this requires them to share a common geometry and may lead to the loss of modality-specific information. In this work, we propose Pesto, a representation learning method that uses a shared topology for alignment while learning a separate embedding space for each modality. We represent the topology as a grid of prototypes, where similar samples are mapped to neighbouring locations. For multimodal data, each modality has its own encoder and prototype matrix, and modalities are aligned through their distributions over the shared map. Thus, different modalities of the same instance are aligned to similar locations on the map, while similar samples are organised into neighbouring regions. We learn the representations and topology jointly by backpropagating the neighbourhood loss through the encoders and using it to update the prototypes. Since alignment takes place on the shared map rather than embedding spaces, each modality can be used separately at inference time without imputation or retraining. Experiments on real-world datasets show that our method retains the characteristics of individual modalities, produces consistent topological maps across modalities, and achieves competitive performance compared with existing methods.

1 INTRODUCTION

A learned representation's quality depends both on individual embeddings and on the global structure of its induced space, which is implicit in most representation learning methods. Encoders map inputs to vectors in $\mathbb{R}^d$, while objectives such as contrastive learning (CL) constrain relationships among selected samples (Chen et al., 2020). The resulting global geometry is typically examined post hoc, through low-dimensional projections that played no role in learning it (van der Maaten & Hinton, 2008; McInnes et al., 2018). This motivated methods that explicitly incorporate a topology into representations (Moor et al., 2020). Self-organising maps (SOMs) achieve this by inducing a grid where nearby cells map to similar samples (Kohonen, 1982). Recent work integrated SOMs with deep encoders, either by using reconstruction-based representation learning (Fortuin et al., 2019) or by incorporating the SOM objective into CL (Huijben et al., 2023). However, those topologies are jointly learned with a single representation space. As many real-world settings are multimodal, methods must capture cross-modal correspondences while enforcing global organisation.

Current multimodal representation learning typically establishes correspondence by mapping modalities into a common embedding space (Radford et al., 2021; Akbari et al., 2021). Modality-specific encoders may process fundamentally different inputs, but their outputs are projected to embeddings of the same dimension and paired observations can be directly compared, aligned through contrastive objectives (Tian et al., 2020), or combined through feature-level fusion (Guo et al., 2019). You et al. (2026) constrain the common-space geometry, aligning multilingual vision-language embeddings with a topological loss so that the two spaces agree beyond paired instances. This makes representations compatible, but ties cross-modal correspondence to embedding parameterisation. Methods with both shared and modality-specific representations are flexible, but separating the two requires further objectives, such as confining the contrastive term to a shared channel and reconstructing what each modality carries on its own (Yates et al., 2026). The same dependence on

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.



(a) One shared map, read by any subset of sensors.

(b) Map trained without labels.

Figure 1: PESTO's shared map and its learned representation for Sleep-EDF-78.

(a) Every modality keeps its own encoder and prototype table $\mathbf{P}_m$ and shares one $\sqrt{K} \times \sqrt{K}$ map. The affinities a_m contract the embedding width d_m away, so the modalities are compared over positions of the map and any subset of them is enough to read it. **(b)** Each cell is coloured by the majority sleep stage of the training samples it wins, with letters at the stage centroids. No label was used to train the map and no classifier is involved, yet 83.2% of adjacent cells share their stage, wakefulness sits far from every sleep stage, and N1, N2 and N3 lie in order of sleep depth.

compatibility appears when only a subset of modalities is available, where existing methods commonly add cross-modal reconstruction or modality dropout to handle missingness (Tsai et al., 2019; Wang et al., 2023). However, cross-modal correspondence can also be defined such that a common topology can relate representations without requiring embeddings to lie in the same space.

Thus we propose PESTO (Projected Embeddings with Self-Trained Organisation), a multimodal representation learning method that learns a shared topology over modality-specific representation spaces. A modality m has its own encoder and set of prototypes, and corresponding prototypes of different modalities are assigned to the same locations on a common two-dimensional grid. Prototype vectors remain modality-specific, while the grid provides a common set of topological coordinates. The similarities of an embedding z_m to the K prototypes define a distribution over grid coordinates, and cross-modal alignment is performed between these distributions rather than between the embeddings, so the alignment objective does not depend on the dimension of each modality. A differentiable SOM objective encourages samples assigned to nearby grid locations to stay close in their respective representation spaces (Kohonen, 1982). Optimising these objectives jointly through the modality encoders learns both how samples from different modalities are aligned and how they are organised over the shared topology, as shown in Figure 1. A classifier head can then read the grid using whichever sensors are present. Figure 1b shows a resulting grid, where unsupervised training for Sleep-EDF-78 (Kemp et al., 2000) organised sleep stages physiologically.

The central empirical question is whether replacing embedding-space alignment with topology alignment preserves representation quality while also keeping meaningful global organisation. Using real-world multimodal datasets, we evaluate representation quality and cross-modal alignment and treat deployment with missing modalities as a consequence of our design rather than as a separate imputation or missingness mechanism. In summary, our contributions are:

- We introduce a multimodal representation learning method that aligns modality-specific representations through a shared topological map; the modalities share grid locations but not prototype vectors, so alignment does not require embeddings of equal dimension.
- We show that the learned map can be read from a single modality without imputation or any change to the model, and that a modality used alone in this way retains substantially more information than a model trained on that modality alone.
- We propagate the SOM objective through the modality encoders, show that this makes the topology shape the representation when several modalities share one map, and evaluate both the topology it produces and its effect on downstream representation quality.

2 Related Work

Deep SOMs learn the input to the SOM grid with an encoder and differ mainly in the objective that shapes it: reconstruction in SOM-VAE (Fortuin et al., 2019) and DESOM (Forest et al., 2021), and contrastive predictive coding in SOM-CPC (Huijben et al., 2023). Irie et al. (2024) instead use the Kohonen rule as the codebook update of a vector-quantised autoencoder, and topological autoencoders reach the same goal without a grid, by matching persistent-homology signatures of the input and latent spaces (Moor et al., 2020). All of these learn one map over one representation space, the best-matching unit (BMU) is selected by Euclidean distance, and the works differ in whether the neighbourhood loss updates the encoder: SOM-VAE stops this gradient, DESOM passes it, and SOM-CPC found no decisive difference between the two. Multimodal ReSOM trains one SOM per modality on fixed features and then connects the maps (Khacef et al., 2020): for every paired sample it takes the BMU of each map, enforces the connection between them, and keeps only the strongest connections, so that one map is read through the other once both are fixed.

VQ-VAE quantises to an unordered codebook (van den Oord et al., 2017), and FSQ replaces the learned codebook with a fixed lattice (Mentzer et al., 2024), whose codes are ordered by construction but not organised by the data. SwAV and SeLa compute assignments over unordered prototypes or clusters and prevent collapse by using Sinkhorn iteration (Caron et al., 2020; Asano et al., 2020; Cuturi, 2013). Cross-modal CL builds on the same kind of code: it treats modalities as views of one underlying content (Tian et al., 2020; Radford et al., 2021), and the choice of repulsion (explicit negatives, redundancy reduction or variance floors) is largely orthogonal to the alignment itself (Chen et al., 2020; Zbontar et al., 2021; Bardes et al., 2022). Both families assume one space in which the comparison is defined, and a prototype in it has an identity but no position.

Methods for missing modalities fall into several groups. Generative methods reconstruct the absent modality from the available ones, with imagination networks (Zhao et al., 2021) or Bayesian meta-learning (Ma et al., 2021). A second group builds one representation that can be computed from any subset of modalities: HeMIS embeds every modality into a common latent space and averages over the modalities that are present (Havaei et al., 2016), mmFormer fuses modality-specific encoders into modality-invariant features (Zhang et al., 2022), factorised models separate shared from modality-specific components (Tsai et al., 2019; Wang et al., 2023), and CPM-Nets learn a latent representation from available views (Zhang et al., 2019). Robust-training methods drop modalities during training so that the fused model degrades gracefully (Ma et al., 2022; Woo et al., 2023), and adaptation methods add small modules that account for the missing input (Lee et al., 2023).

PESTO differs from these works as follows. All modalities share one map, each keeping its own encoder and its own prototypes, being aligned through a position on the grid and not a vector, and the encoders are learned with the map rather than associated once both are fixed, as in ReSOM. The prototypes carry grid coordinates and the neighbourhood term arranges them there while the encoders are trained, so the comparison between modalities is defined over positions on a map rather than over an unordered set of prototypes in one embedding space. A subset of the modalities is read by averaging their distributions over that same map, so that nothing is imputed and nothing is retrained, where HeMIS averages embeddings in one latent space common to all of them.

3 Method

3.1 Problem Formulation

Conceptually, our framework seeks to map heterogeneous input modalities into a unified, low-dimensional topological manifold. Rather than relying on simple vector metric spaces standard in contrastive learning, we construct a shared discrete grid, give each modality its own encoder and its own prototype table on it, and align the modalities through their distributions over its cells (Figure 1a). We consider multimodal representation learning with M synchronously recorded data streams. Given paired observations $(x_1, \ldots, x_M)$, where x_m denotes the input of modality m and has its own shape, the objective is to learn encoders that map all modalities into a shared discrete representation space. The representation is organised as a SOM prototype grid, which imposes topological structure on the learned embedding spaces. No label supervision is used during training, and alignment is achieved purely through the co-occurrence of the modalities.

3.2 PER-MODALITY TABLES ON A SHARED MAP

The map is a square grid of K cells. A modality $m \in \{1, \ldots, M\}$ has an encoder that outputs an embedding $z_m \in \mathbb{R}^{d_m}$, and a prototype table $\mathbf{P}_m \in \mathbb{R}^{K \times d_m}$ whose k-th row represents cell k for that modality. A sample's affinity to the map is its cosine similarity to every prototype,

$$a_m \;=\; \widehat{z}_m \, \widehat{\mathbf{P}}_m^{\top} \in \mathbb{R}^K, \tag{1}$$

where $\widehat{\cdot}$ denotes ℓ_2 normalisation; the BMU of modality m is $\hat{k}_m = \arg\max_k a_m[k]$. Normalising both the embedding and the prototypes removes the dependence on the magnitude of z_m. Cell assignments thus depend only on direction, allowing encoders with different output norms to use the same map. The modalities share a common map rather than a common embedding space. The affinity computation maps each modality-specific embedding of dimension d_m to a K-dimensional vector over the shared cells. Prototype tables with different widths therefore support the same cross-modal objective as a single shared table. Modalities agree when they select the same map location, while each retains its own embedding geometry and output scale.

3.3 TRAINING OBJECTIVE

Two terms train the encoders and the tables jointly. Let $\pi_m = \mathrm{softmax}(a_m/\tau)$ be the activation map of a sample under modality m, with map temperature τ. Different modalities of the same sample should activate the same region of the grid, and different samples should not. We contrast the activation maps rather than the embeddings, over every ordered pair of modalities:

$$\mathcal{L}_{\text{align}} \;=\; \sum_{m \neq m'} \frac{1}{B} \sum_{i=1}^{B} \mathrm{CE}\big(s_{i\cdot}^{mm'},\, i\big), \tag{2}$$

where $\mathrm{CE}(s_{i\cdot}^{mm'}, i)$ is the cross-entropy of identifying sample i among the B samples of the batch from its similarities $s_{i1}^{mm'}, \ldots, s_{iB}^{mm'}$. The map of each sample under one modality is pulled toward its map under another and pushed away from the maps of the other samples. This is the only term that separates samples, and it is what prevents collapse (Appendix G). Because it is computed on activation maps, it enforces grid-wise agreement.

A grid becomes a map when neighbouring cells represent similar samples. Each sample therefore pulls its BMU and, with decaying strength, the cells around it. Let $c = \arg\max_k \frac{1}{M} \sum_m a_m[k]$ be the consensus cell of a sample, $g_k \in \mathbb{Z}^2$ the row and column of cell k on the grid, and $h_k(c) = \exp(-\|g_c - g_k\|^2 / 2\sigma^2)$ a Gaussian over grid positions whose width σ is measured in cells. Then

$$\mathcal{L}_{\text{grid}} \;=\; \sum_{m} \frac{1}{d_m} \sum_{k=1}^{K} h_k(c) \, \big\| \mathbf{P}_m[k] - \widehat{z}_m \big\|_2^2, \tag{3}$$

averaged over the batch, with σ annealed linearly from 3 to 0.5 cells during training. The target is the normalised embedding $\widehat{z}_m$, the same vector that selects the BMU in Equation 1, so the term cannot be reduced by changing the embedding's norm. The training objective is $\mathcal{L} \;=\; \mathcal{L}_{\text{align}} + \lambda \, \mathcal{L}_{\text{grid}}$ (Appendix C has λ values per benchmark). The gradient of $\mathcal{L}$ reaches the encoder, pressuring the embeddings to help neighbouring cells map to similar samples (see Section 4.4).

3.4 READING THE MAP

After unsupervised training, a classifier head reads the map through the affinities of the modalities that are present. For a set $S \subseteq \{1, \ldots, M\}$ of observed modalities, the head input is

$$\bar{\pi}_S \;=\; \mathrm{softmax}\Big(\frac{1}{|S|} \sum_{m \in S} a_m \,/\, \tau_r \Big), \tag{4}$$

with readout temperature τ_r. The head is trained with all modalities and deployed on any observed subset S: the same equation applies, and its input stays a distribution over the same K cells, so nothing is retrained or imputed and absent encoders are never executed. Fusion by concatenation or attention lacks this property, as its representation exists only after every encoder runs. The head

is trained using one of two protocols, which differ in whether the prototype tables move with it. The frozen protocol trains the head alone and reads the unsupervised representation as it stands; `proto_head` trains the tables together with the head, and is the model we deploy. The encoders are fixed in both. The grid can also be analysed without a learned head, by labelling cells with the majority class of the training samples they win (direct readout, `acc_strict`; Appendix A).

4 EXPERIMENTS

We evaluate PESTO on three benchmarks whose sensors differ in kind, number and failure mode: a wrist-worn inertial unit and a skeleton tracker (UTD-MHAD, Chen et al., 2015), scalp EEG, EOG and chin EMG recorded overnight (Sleep-EDF, Kemp et al., 2000; Goldberger et al., 2000), and depth cameras, mmWave radar and body keypoints for activity recognition by unseen subjects (MM-Fi, Yang et al., 2023b). The experiments answer three questions. How well does a map shared by modalities that keep their own embedding spaces perform on each benchmark, and how does it compare with the state of the art under the established protocol of that benchmark (§4.2)? What does the shared map provide when sensors are missing at test time (§4.3)? What do the grid and its topology cost and contribute (§4.4)?

4.1 EXPERIMENTAL SETUP

Table 1: Evaluation protocols and corresponding comparisons.

Benchmark	Unsupervised inputs	Size / classes	Test split	Reference protocol	Results
UTD-MHAD	IMU, skeleton	861 / 27	instances, 5 splits	internal control	Tab. 5
UTD-MHAD	IMU, skeleton	861 / 27	subjects, odd/even	NFC, HNS	Tab. 6
Sleep-EDF-78	EEG, EOG	192,233 / 5	subjects, 10-fold	XSleepNet	Tab. 4
Sleep-EDF-78	EEGs, EOG, EMG	195,479 / 5	subjects, 5-fold	SleepSMC	Tab. 8
Sleep-EDF-20	EEGs, EOG, EMG	42,308 / 5	subjects, fixed	CroSSL	Tab. 3
Sleep-EDF-20	EEG, EOG	42,308 / 5	subjects, LOSO	MultiCons	Tab. 4
MM-Fi	keypoints, depth, radar	560 / 14	subjects, fixed	MMBind	Tab. 7

Sleep-EDF-78 is the complete Sleep Cassette study (78 subjects, 153 nights) and Sleep-EDF-20 its first 20 subjects (39 nights), the subset prior self-supervised studies use. A sample is an action instance on UTD-MHAD, a 30-second epoch on Sleep-EDF and a 30-frame sequence on MM-Fi. MM-Fi keypoints are used during unsupervised training only. Appendix B specifies each protocol in full.

Table 1 summarises the seven comparative settings. The comparators are the works that established these benchmarks: contrastive pre-training on wearable and skeletal streams, with hard negatives (HNS, Choi et al., 2023) or without them (NFC, a name we give to the unnamed model of Yang et al., 2023a); multimodal self-supervised sleep staging (Deldari et al., 2022; 2024; Yu et al., 2025); training with multiple sensors for single-sensor deployment (Ma et al., 2025); supervised sleep sequence models (Supratak et al., 2017; Phan et al., 2019; 2022); and methods that connect sensors through a shared modality (Ouyang et al., 2025). For each comparison, we follow the data preprocessing, data splits, model-selection rule, and evaluation metric of the corresponding work. We reconstructed these protocols from the authors' released code and verified the resulting datasets against their reported statistics before training (Appendix B). Consequently, the same benchmark may appear under multiple settings when prior studies adopt different evaluation protocols, as is particularly the case for Sleep-EDF (Kemp et al., 2000; Goldberger et al., 2000).

PESTO is trained without labels: they are used neither for optimisation nor for model selection, and all configurations are selected on validation data. We then train a classifier using either the frozen or the `proto_head` protocol described in Section 3.4; neither updates the encoders, and both read the map using Equation 4, which Appendix F compares with reading the affinities directly. The detailed tables identify comparators that fine-tune their encoders or use labels during representation learning. We report these methods under their published protocols and additionally match two conditions to isolate specific factors: label-informed negatives on UTD-MHAD (Table 6) and encoder fine-tuning on MM-Fi (Table 7). For each comparison, we use the encoder architecture of the corresponding work and train it from scratch, controlling for differences in backbone capacity: DeepSleepNet (Supratak et al., 2017) for sleep, MMBind for MM-Fi, and a 1D CNN with HCN (Li et al., 2018) for UTD-MHAD. For sequence-level sleep staging, we additionally use the recurrent block of DeepSleepNet over 25 consecutive epochs.

Table 2: Summary of all seven comparisons, grouped by comparison setting.

Setting	Baseline training	PESTO training	Test input	Metric	Baseline → PESTO	Details
Against self-supervised methods: no labels before the classifier on either side						
Sleep-EDF-20 / CroSSL	self-supervised, frozen	unsupervised, frozen	all four	F1	$0.722 \rightarrow 0.785_{\pm 0.004}$	Tab. 3
Sleep-EDF-20 / MultiCons	self-supervised, fine-tuned	unsupervised, `proto_head`	EEG, EOG	Acc. / MF1	$85.8/79.2 \rightarrow 85.9/79.5$	Tab. 4
UTD-MHAD, subjects	NFC, self-supervised	unsupervised, `proto_head`	IMU, skeleton	Acc.	$96.06 \rightarrow 96.2_{\pm 0.7}$	Tab. 6
Against supervised models trained end to end with labels						
Sleep-EDF-78 / XSleepNet	supervised, end to end	unsupervised, `proto_head`	EEG, EOG	Acc. / MF1 / κ	$84.0/78.7/.778 \rightarrow 84.3/78.0/.781$	Tab. 4
Sleep-EDF-78 / SleepSMC	supervised, end to end	unsupervised, `proto_head`	all four	Acc.	$81.6 \rightarrow 83.9$	Tab. 8
Comparisons we control: same encoders, inputs and classifier, only the objective or the sensor set differs						
UTD-MHAD, instances	one sensor, `proto_head`	both sensors, `proto_head`	IMU	Acc.	$83.4_{\pm 1.8} \rightarrow 92.9_{\pm 2.8}$	Tab. 5
MM-Fi / MMBind	contrastive, frozen	unsupervised, frozen	depth, radar	Acc.	$74.3_{\pm 5.6} \rightarrow 81.0_{\pm 2.9}$	Tab. 7

Arrows compare the baseline with PESTO within one row. The training columns refer only to representation learning, a task classifier is then trained. `proto_head` updates our prototype tables together with that classifier and leaves the encoders frozen. Subscripts denote standard deviations over five random seeds. MF1 is the macro-averaged F1 score. Except for the CroSSL F1, values are percentages.

Table 3: Sleep-EDF-20 under the self-supervised protocol.

Encoders	Sensors	COCOA	CroSSL	PESTO
frozen	all	0.628	0.722	**0.785** $\pm$.004
	missing at test	—	0.694	**0.729** $\pm$.005
	missing in training and test	—	0.667	**0.735** $\pm$.004
fine-tuned / `proto_head`	all	0.678	0.741	**0.791** $\pm$.003
	missing at test	—	**0.739**	0.728 $\pm$.002
	missing in training and test	—	0.581	**0.735** $\pm$.003

F1 as computed by the CroSSL code; PESTO: mean $\pm$ s.d. over 5 seeds. "Missing" averages over all 15 subsets of the four sensors, including the full set. COCOA and CroSSL are quoted from Deldari et al. (2024), where a supervised end-to-end model scores 0.717 with all sensors.

The objective formulation is shared across benchmarks and negative sampling follows the convention of each domain. For sleep, the alignment term draws negatives from the same recording, following common practice in self-supervised sleep staging (Banville et al., 2021; Huijben et al., 2023). For MM-Fi, negatives are drawn from the same subject: each batch contains all 14 actions of one subject, keeping subject identity and environment fixed within the batch. MM-Fi features are temporally averaged before entering the map; keypoints are expressed in the body frame, and depth is represented relative to its static background. The contrastive baseline in Table 7 uses the MMBind contrastive objective with the same encoders, inputs, batches, and classifier architecture. Appendix C provides the remaining implementation details.

4.2 MAIN RESULTS

Table 2 summarises every comparison in the paper, grouped by what PESTO is compared with rather than by dataset, as that is what the seven protocols have in common. Against other self-supervised methods, PESTO wins on every benchmark, while matching or exceeding fully supervised models. In comparisons where encoders, inputs, batches and classifier are held fixed and only the objective or the sensor set changes, the margin is largest, placing the gain in the map rather than in the backbone or the data pipeline. Each row compares two systems under one protocol. The three Sleep-EDF protocols use different dataset subsets, sensor sets and subject splits. We therefore report them separately: CroSSL tests frozen self-supervised representations and missing sensors, MultiConsSleepNet fine-tunes its encoders under leave-one-subject-out evaluation, and XSleepNet trains supervised sequence models end to end on the larger Sleep-EDF-78 cohort.

Table 3 follows the protocol of CroSSL, including its three test conditions. With frozen encoders, PESTO outperforms the best self-supervised representation in every condition. With all sensors, the frozen map is also more accurate than CroSSL with its encoders fine-tuned end to end and than a supervised model trained end to end (0.717), although its encoders never see a label. With `proto_head`, which trains the prototype tables and leaves the encoders frozen, PESTO maintains its strong performance, only being outperformed by CroSSL by a small margin in the missing at test scenario. Further results (Appendix F) show that with one sensor removed at test time, and even with two, the frozen map remains more accurate than fine-tuned CroSSL with all four sensors.

The frozen, unsupervised encoders also hold their own against fully supervised sequence models. Under XSleepNet's protocol, PESTO outperforms both XSleepNet and SeqSleepNet on accuracy and

Table 4: Comparison with published sleep-staging protocols on Sleep-EDF.

Method	Training	Test input	Acc.↑	MF1↑	κ↑
Sleep-EDF-78, 10-fold subject-wise cross-validation (XSleepNet)					
SeqSleepNet	supervised, end to end	EEG, EOG	83.8	78.2	0.776
XSleepNet	supervised, end to end	EEG, EOG	84.0	**78.7**	0.778
XSleepNet, trained on EEG only	supervised, end to end	EEG	84.0	77.9	0.778
PESTO (`proto_head`)	unsupervised, frozen	EEG, EOG	**84.3**	78.0	**0.781**
Sleep-EDF-20, 20-fold leave-one-subject-out (MultiConsSleepNet)					
MultiConsSleepNet	self-supervised, fine-tuned	EEG, EOG	85.8	79.2	—
PESTO (`proto_head`)	unsupervised, frozen	EEG, EOG	**85.9**	**79.5**	**0.806**

Metrics are pooled over folds in both blocks. The upper block uses the folds of Phan et al. (2022) and quotes all three of its rows from that paper, which reports the same accuracy and κ for its EEG-only model as for its EEG and EOG model. The lower block quotes MultiConsSleepNet from Yu et al. (2025) after end-to-end fine-tuning. Our encoders are never updated in either block.

Table 5: UTD-MHAD: test accuracy (%), mean ± s.d. over five stratified splits.

Unsupervised training	Test input	frozen	`proto_head`
PESTO, IMU and skeleton	IMU, skeleton	96.6 ±1.6	**98.1** ±0.7
	skeleton	94.9 ±2.2	95.8 ±2.6
	IMU	89.9 ±3.8	92.9 ±2.8
IMU, two augmented views	IMU	80.0 ±3.2	83.4 ±1.8

κ; XSleepNet keeps the higher macro-F1 (Table 4). Table 4 also shows MultiConsSleepNet, a recent contrastive method that aligns EEG and EOG before fine-tuning, under its leave-one-subject-out protocol. MultiConsSleepNet fine-tunes its encoders end to end; PESTO is ahead on both accuracy and macro-F1 with its encoders frozen.

We evaluate UTD-MHAD in two ways. Five stratified instance splits isolate the effect of multimodal rather than single-modal representation learning, whereas the odd/even-subject split compares with published methods on entirely unseen subjects. On UTD-MHAD the unsupervised map recognises 98.1% of test actions with `proto_head` and 96.6% when only the head is trained (Table 5). The skeleton alone keeps 95.8% and the IMU alone 92.9%. Single-sensor accuracy is itself a product of multimodal training. The same architecture trained on two augmented views of the IMU reaches 83.4%, so adding the skeleton for unsupervised training raises IMU-only accuracy by 9.5 points on every split (§4.3). The map itself also works as a classifier: labelling each cell by the majority class of the training instances it wins achieves 87.4% test accuracy, against 3.7% for chance.

Table 6 repeats the comparison across subjects, where every test subject is unseen. Without a label before the classifier, the map reaches $96.2\pm0.7\%$, ahead of the strongest self-supervised comparator, NFC (96.06%), and it serves either sensor alone: 93.6% from the skeleton, 7.6 points above NFC (86.05%), and 71.1% from the inertial stream. Choi et al. (2023) instead select the negatives with the labels, keeping only samples whose class differs from the anchor. The label-free map is already 4.1 points above HNS (92.06%) and above their supervised multimodal model (92.04%); with those negatives and the augmentations of that work it reaches 96.9%, 4.8 points above HNS.

MM-Fi provides a third modality combination and a fixed unseen-subject protocol. The matched contrastive control uses the same encoders, inputs, batches and classifier as PESTO, making it the most direct test of the map objective; the published MMBind result provides an additional reference but uses a different training pipeline. Table 7 reports MM-Fi under the protocol of MMBind, where every test subject is unseen during both unsupervised and supervised training. With frozen encoders, PESTO achieves 81.0% test accuracy, is more accurate and more stable than the matched contrastive control, and outperforms MMBind with supervised encoder fine-tuning. It also remains far more accurate when a sensor is absent both for depth and radar alone.

4.3 DEPLOYMENT WITH MISSING SENSORS

We evaluate missing-modality deployment across complementary settings. On Sleep-EDF, we consider all sensor subsets as well as single-sensor deployment, allowing direct comparison with pub-

Table 6: UTD-MHAD across subjects: test accuracy (%).

Pre-training	Labels used	IMU, skeleton	skeleton	IMU
NFC	none	96.06	86.05	75.58
CMC	none	95.12	—	—
Barlow Twins	none	91.63	—	—
Supervised, multimodal	all	96.51	—	—
HNS	negatives	92.06	—	—
CMC	negatives	91.84	—	—
Supervised, multimodal	all	92.04	—	—
PESTO	none	**96.2** ±0.7	**93.6** ±2.3	71.1 ±1.8
PESTO, their negatives	negatives	**96.9** ±0.8	**95.1** ±1.1	**72.0** ±2.0

The odd/even subject split of both compared works. Baselines are quoted as published, the upper block from Yang et al. (2023a) and the middle one from Choi et al. (2023); ours are means ± s.d. over five seeds with `proto_head`. Our second row adopts the label-informed negatives of the middle block. Each of our rows takes the augmentations of the work it faces, and only the first blurs the maps over grid positions in the alignment term (σ=2; Appendix C).

Table 7: MM-Fi under the protocol of MMBind: accuracy and macro-F1 (%).

Method	Encoders	Accuracy	Macro-F1	Depth	Radar
MMBind upper bound	fine-tuned	72.5	72.0	–	–
Contrastive pre-training	frozen	74.3 ±5.6	73.6 ±6.2	60.1 ±1.9	28.3 ±4.7
PESTO	frozen	**81.0** ±2.9	**81.1** ±2.9	**75.2** ±0.8	**58.7** ±5.3
PESTO	fine-tuned	**81.1** ±2.3	**81.0** ±2.0	**73.6** ±4.0	**47.9** ±4.9

Activity recognition on 16 unseen subjects from depth and mmWave radar, and accuracy when only one of the two is present at test time; mean ± s.d. over 5 seeds. Contrastive pre-training is the matched baseline of §4.1. The MMBind model pre-trained on the same naturally paired data (its upper bound) is quoted from Ouyang et al. (2025) and fine-tunes its encoders with labels.

lished missing-sensor results. On MM-Fi and UTD-MHAD, we evaluate reduced-sensor deployment using radar alone and IMU alone, respectively. Together, these settings cover both sensor loss and deliberate sensor reduction without retraining the multimodal model. We additionally report EEG-only inference on Sleep-EDF as the two-modality instance of the same mechanism.

PESTO is designed so the affinities of any subset of sensors still form a distribution over the whole map, which a classifier reads. Accuracy decreases gradually as sensors are removed (Figure 3): with frozen encoders it is 0.785 with all four sensors, 0.773 with three and 0.750 with two, and with `proto_head` 0.791, 0.776 and 0.746. When sensors are missing at test time, the frozen map is more accurate than CroSSL (Table 3). PESTO needs no preparation for this: training with randomly removed sensors leaves it where it was (0.729 and 0.735). Fine-tuned CroSSL falls to 0.581 when sensors are missing during training as well, while the frozen map stays at 0.735. The same holds on MM-Fi, with a different pair of sensors (Table 7). The classifier is trained on depth and radar and then read with one of them removed: the frozen map recognises 75.2% of the actions from depth alone and 58.7% from radar alone, against 60.1% and 28.3% for contrastive pre-training.

SleepSMC trains on all sensors with labels and is designed to be deployed on one. Table 8 compares the two under its protocol. The frozen map is ahead in every deployment. One unsupervised PESTO model serves all four deployments. Under XSleepNet's protocol, dropping EOG at test time costs PESTO 1.1 accuracy points. On UTD-MHAD, PESTO trained on both streams keeps 95.8% with skeletons alone and 92.9% with IMUs alone, against 98.1% with both (Table 5). Multimodal training also improves what sensors can do alone. On UTD-MHAD, a map trained on both streams and deployed on the inertial stream is more accurate than the same architecture trained on two augmented views of that stream alone, on all five splits (Table 5, Figure 4a). The second stream improves the map itself: direct readout rises from 0.620 to 0.874 (Appendix G).

4.4 ANALYSIS

Figures 1b and 2 label cells with a majority vote of the training samples assigned to them. The figures come from different protocols, two sensors over all 78 subjects and four sensors over the 20-subject subset, so what they share is a property of the map rather than of one sensor set or one split. Sleep

Table 8: Training on all sensors and deploying on fewer, Sleep-EDF-78.

Method	Training	Test input	Acc.↑	MF1↑	κ↑
Four sensors, 5-fold subject-wise cross-validation (SleepSMC)					
SleepSMC	supervised	all four	81.6	75.6	0.745
		EEG only	79.6	72.2	0.717
		EOG only	73.1	64.9	0.626
		EMG only	52.9	31.7	0.302
PESTO (`proto_head`)	unsupervised, frozen	all four	**83.9**	**77.9**	**0.777**
		EEG only	**83.5**	**77.0**	**0.770**
		EOG only	**79.9**	**73.0**	**0.721**
		EMG only	**56.1**	**38.4**	**0.356**
Two sensors, 10-fold subject-wise cross-validation (XSleepNet)					
PESTO (`proto_head`)	unsupervised, frozen	EEG, EOG	84.3	78.0	0.781
		EEG only	83.2	76.5	0.766

One model per block, trained once on all the sensors of that protocol and then read with the sensors listed, with nothing retrained and no imputation. SleepSMC is supervised and its numbers are quoted from Ma et al. (2025); it is designed for this case. The first row of the lower block is the model of Table 4, repeated as the reference point for the removal below it.

stages form contiguous regions which follow sleep physiology: 83.2% of adjacent cells share their stage, against $32.5 \pm 1.6\%$ if labels were random (79.3% against $24.3 \pm 1.5\%$ on Sleep-EDF-20). Wakefulness lies 6 to 12 cells from every sleep stage. On Sleep-EDF-78, N1, N2 and N3 lie roughly along a line in order of depth, and N1 lies next to REM, whose EEG it closely resembles. Using these majority-vote labels, the map alone achieves 75.7% test accuracy (76.0% on Sleep-EDF-20), and 74.8% from the EEG alone. Sensors also agree on the grid: EEG and EOG select the same cell 24 to 113 times more often than unpaired samples do, the chin EMG about 5 times (Appendix D).

Across 24 paired configurations on UTD-MHAD, per-sensor prototype tables and one shared table differ by 0.06 points under `proto_head` ($t{=}-0.25$) and 1.4 points under frozen ($t{=}-1.45$); neither was significant. Only per-sensor tables admit encoders of different widths, which MM-Fi needs: its keypoint encoder produces vectors with $d_m = 16$ and its depth and radar encoders $d_m = 64$.

Larger grids help slightly and then saturate. On a validation fold of Sleep-EDF-78, macro-F1 rises from 76.2 at $K{=}100$ to 77.3 at $K{=}900$, and $K{=}1600$ adds 0.1; on UTD-MHAD, the four grid sizes we tried lie within one standard error of each other. Appendix E reports every value.

When the neighbourhood term stops at the encoder, as in SOM-VAE, the grid is fitted over a representation decided by the alignment term alone. Letting its gradient reach the encoder improves the layout of the map. The correlation between grid distance and class dissimilarity as a fraction of its attainable maximum increases on UTD-MHAD at every grid size with capacity to spare, on all five splits (Figure 4b, Table 10). This one switch moves the layout as far as the entire range of grid sizes does, at a small cost to direct grid readout at $K{=}400$. Gradient norms explain why the route matters. On the prototype tables the alignment term is 60 times stronger than the neighbourhood term, so the topology acts on the map mainly through the encoder (Appendix H).

5 CONCLUSION

We introduced PESTO, which aligns modalities through a topological map that they learn together, instead of through a common embedding space. Each modality keeps its own encoder and its own representation; the shared map relates them to one another and places similar samples in neighbouring regions. Because every modality reaches the map on its own, a model trained on all sensors can be used with any subset of them, without imputation or retraining. On sleep staging and activity recognition, the map learned without labels outperforms multimodal self-supervised methods, matches a supervised sleep-staging model, and stays accurate when sensors are missing. Learning with a second modality also improves what a single sensor achieves on its own, and the map organises the data in a meaningful way, with sleep stages forming contiguous regions found without labels. A shared topology is thus a practical alternative to a shared embedding space for multimodal representation learning.

AI USE STATEMENT

We used generative AI tools to polish the wording of this paper.

ETHICS STATEMENT

All experiments use public datasets recorded from consenting volunteers and released in de-identified form: Sleep-EDF from PhysioNet, UTD-MHAD and MM-Fi. No new data were collected, and no attempt was made to identify any participant.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendices B and C give the protocols and all settings. Each data pipeline reproduces the sample counts of the work it follows. Every number in the tables is re-aggregated from the released per-run result files by a script that asserts against the printed values. The code will be made publicly available upon publication of the paper.

REFERENCES

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24206–24221, 2021.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021. doi: 10.1088/1741-2552/abca18.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2105.04906.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9912–9924, 2020.

Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *IEEE International Conference on Image Processing (ICIP)*, pp. 168–172, 2015. doi: 10.1109/ICIP.2015.7350781.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Hyeongju Choi, Apoorva Beedu, and Irfan Essa. Multimodal contrastive learning with hard negative sampling for human activity recognition. *arXiv preprint arXiv:2309.01262*, 2023. doi: 10.48550/arXiv.2309.01262.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pp. 2292–2300, 2013.

Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V. Smith, and Flora D. Salim. COCOA: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–28, 2022. doi: 10.1145/3550316.

Shohreh Deldari, Dimitris Spathis, Mohammad Malekzadeh, Fahim Kawsar, Flora D. Salim, and Akhil Mathur. CroSSL: Cross-modal self-supervised learning for time-series through latent masking. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 152–160, 2024. doi: 10.1145/3616855.3635795.

Florent Forest, Mustapha Lebbah, Hanene Azzag, and Jérôme Lacaille. Deep embedded self-organizing maps for joint representation learning and topology-preserving clustering. *Neural Computing and Applications*, 33:17439–17469, 2021. doi: 10.1007/s00521-021-06331-w.

Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. SOM-VAE: Interpretable discrete representation learning on time series. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1806.02199.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.cir.101.23.e215.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. doi: 10.1109/ACCESS.2019.2916887.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, volume 9901 of *Lecture Notes in Computer Science*, pp. 469–477. Springer, 2016. doi: 10.1007/978-3-319-46723-8_54.

Iris A. M. Huijben, Arthur Andreas Nijdam, Sebastiaan Overeem, Merel M. Van Gilst, and Ruud J. G. Van Sloun. SOM-CPC: Unsupervised contrastive learning with self-organizing maps for structured representations of high-rate time series. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14132–14152. PMLR, 2023. arXiv:2205.15875.

Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. Self-organising neural discrete representation learning à la kohonen. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, pp. 343–362. Springer, 2024. doi: 10.1007/978-3-031-72332-2_23.

B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Lyes Khacef, Laurent Rodriguez, and Benoît Miramond. Brain-inspired self-organization with cellular neuromorphic computing for multimodal unsupervised learning. *Electronics*, 9(10):1605, 2020. doi: 10.3390/electronics9101605.

Kimmo Kiviluoto. Topology preservation in self-organizing maps. In *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, pp. 294–299, 1996. doi: 10.1109/ICNN.1996.548907.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982. doi: 10.1007/BF00337288.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952, 2023. doi: 10.1109/CVPR52729.2023.01435.

Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 786–792, 2018. doi: 10.24963/ijcai.2018/109.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, 2022. doi: 10.1109/CVPR52688.2022.01764.

Shuo Ma, Yingwei Zhang, Yiqiang Chen, Hualei Wang, Yuan Jin, Wei Zhang, and Ziyu Jia. SleepSMC: Ubiquitous sleep staging via supervised multimodal coordination. In *International Conference on Learning Representations (ICLR)*, 2025.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. doi: 10.48550/arXiv.1802.03426.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations (ICLR)*, 2024.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7045–7054. PMLR, 2020. arXiv:1906.00722.

Xiaomin Ouyang, Jason Wu, Tomoyoshi Kimura, Yihan Lin, Gunjan Verma, Tarek Abdelzaher, and Mani Srivastava. MMBind: Unleashing the potential of distributed and heterogeneous data for multimodal learning in IoT. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pp. 491–503. ACM, 2025. doi: 10.1145/3715014.3722053.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019. doi: 10.1109/TNSRE.2019.2896659.

Huy Phan, Oliver Y. Chen, Minh C. Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2022. doi: 10.1109/TPAMI.2021.3070057.

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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. arXiv:2103.00020.

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017. doi: 10.1109/tnsre.2017.2721116.

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. Springer, 2020. doi: 10.1007/978-3-030-58621-8_45.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1806.06176.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 6306–6315, 2017. arXiv:1711.00937.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887, 2023. doi: 10.1109/CVPR52729.2023.01524.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2776–2784, 2023. doi: 10.1609/aaai.v37i3.25378.

Huaigang Yang, Ziliang Ren, Huaqiang Yuan, Zhenyu Xu, and Jun Zhou. Contrastive self-supervised representation learning without negative samples for multimodal human action recognition. *Frontiers in Neuroscience*, 17:1225312, 2023a. doi: 10.3389/fnins.2023.1225312.

Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. MM-Fi: Multi-modal non-intrusive 4D human dataset for versatile wireless sensing. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, 2023b. doi: 10.52202/075280-0822.

Don J. Yates, Hakki Erhan Sevil, and Arash Mahyari. Disentangling shared and specific features in deep multimodal clustering with contrastive-complementary learning. *Pattern Recognition Letters*, 207:274–280, 2026. doi: 10.1016/j.patrec.2026.07.008.

Junwon You, Dasol Kang, and Jae-Hun Jung. Topological alignment of shared vision-language embedding space. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1207–1215. PMLR, 2026.

Yangzuyi Yu, Shuyu Chen, Wanxin Wei, Jingcong Li, and Jiahui Pan. MultiConsSleepNet: Self-supervised contrastive learning for a multimodal consistency-based automatic sleep staging model. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7, 2025. doi: 10.1109/EMBC58623.2025.11251569.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. arXiv:2103.03230.

Changqing Zhang, Zongbo Han, Yajie Cui, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. CPM-Nets: Cross partial multi-view networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 557–567, 2019.

Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2022*, volume 13435 of *Lecture Notes in Computer Science*, pp. 107–117. Springer, 2022. doi: 10.1007/978-3-031-16443-9_11.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2608–2618. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.203.

6 APPENDIX

A METRIC DEFINITIONS

`acc_strict` (micro; direct grid readout). Each cell k is labelled with the majority class $y(k)$ of the training samples whose consensus cell it is, and a test sample i is scored by the label of its consensus

cell $c^{(i)}$,

$$\texttt{acc_strict} \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big[y(c^{(i)}) = y^{(i)}\Big]\,. \tag{5}$$

Cells never activated in training count as errors. Reading through one modality replaces $c^{(i)}$ by $\hat{k}_m^{(i)}$. Random $= 1/27 = 0.037$ on UTD-MHAD.

cross-modal agreement. The fraction of test samples on which two modalities select the same cell,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big[\hat{k}_m^{(i)} = \hat{k}_{m'}^{(i)}\Big]\,, \qquad \hat{k}_m^{(i)} = \arg\max_k a_m^{(i)}[k], \tag{6}$$

together with the fraction for which the two cells lie at most one step apart on the grid, under Chebyshev distance. Both are reported against the same statistic computed after randomly permuting the assignments of one modality, which keeps the distribution of each modality over the cells and removes only the pairing. Random selection would give $1/K = 0.25\%$ at $K = 400$.

`layout`. With $G[a, b]$ the mean grid distance between the BMUs of class a's and class b's samples and $D[a, b] = 1-\cos(\mu_a, \mu_b)$ the dissimilarity of their encoder-space centroids, layout is $\mathrm{corr}(G, D)$ over class pairs, as a percentage of the ceiling attainable by optimally re-placing the 27 classes on the grid of that run, found by local search. The ceiling is computed per run, since it depends on the class geometry of that run. Random assignment yields a mean score of 0.000 ± 0.050 over 300 trials. The estimated ceiling for a 10×10 grid ranges from $+0.71$ to $+0.83$, compared with $+0.59$ for a one-dimensional chain, indicating that the addition of a second grid dimension provides a measurable performance advantage.

Prototype cosine. Mean pairwise cosine between rows of a prototype table, maximised over tables; > 0.9 indicates collapse. It is the most sensitive collapse detector we found: a configuration can reach 0.995 while still scoring 0.67 on direct grid readout, so neither accuracy nor readout alone detects it.

Adjacent cells with the same stage. With $y(k)$ as above and E the set of pairs of labelled cells one step apart horizontally or vertically,

$$\mathrm{adj} \;=\; \frac{1}{|E|}\sum_{(k,k')\in E} \mathbb{1}[y(k) = y(k')]\,. \tag{7}$$

The random reference permutes the labels over the labelled cells, which keeps the number of cells per stage and destroys only their placement; we report its mean and standard deviation over 1000 permutations.

TE (Kiviluoto, 1996). With $k_1^{(i)}$ and $k_2^{(i)}$ the cells of highest and second-highest affinity of sample i under one modality,

$$\mathrm{TE} \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big[\big\|g_{k_1^{(i)}} - g_{k_2^{(i)}}\big\|_\infty > 1\Big]\,, \tag{8}$$

the fraction of samples whose two best cells are not among the eight grid neighbours of one another; a random pair of cells gives $1 - 8/K$.

Quantisation error is not reported: 94.6% of the QE of the skeletal stream is norm mismatch, which a cosine BMU cannot see, and QE is exactly what a commitment term minimises, making it a training objective rather than an independent measurement.

B BENCHMARK PROTOCOLS

Each pipeline was re-implemented from the released code of the compared work and checked against its reported sample counts before any model was trained. Where a paper and its code disagree, we follow the code and say so.

Sleep-EDF-78, supervised protocol (XSleepNet). The Sleep Cassette study contains 153 nights from 78 subjects (Kemp et al., 2000). Following the XSleepNet code, each night is cropped from 30 minutes before lights-off to 30 minutes after lights-on, band-pass filtered at 0.3–40 Hz, and kept in microvolts without normalisation; stages 3 and 4 are merged, and movement and unscored epochs are removed. This yields 192,233 epochs, with the count of every night identical to the file list of XSleepNet. We use the Fpz-Cz EEG and horizontal EOG channels and the ten folds of the split file released with XSleepNet, including its validation subjects. As in XSleepNet, the checkpoint with the highest validation accuracy is kept, and metrics are computed on the confusion matrix summed over the ten test folds.

Sleep-EDF-20, self-supervised protocol (COCOA, CroSSL). COCOA states that its preprocessing follows DeepSleepNet; we therefore ported the DeepSleepNet preparation script and extended it to the four sensors of CroSSL (Fpz-Cz and Pz-Oz EEG, horizontal EOG, submental EMG, all at 100 Hz). The first 20 subjects give 39 nights and 42,308 epochs, the count reported by DeepSleepNet. CroSSL oversamples minority classes in the training and validation sets, which gives 55,414 epochs and matches the "55K" of its table. The split follows the CroSSL loader: with recordings sorted by name, the last eight nights (subjects 16–19) form the test set and the seven before them (subjects 12–15) the validation set. The evaluation classifier is that of CroSSL (a 128-unit hidden layer with dropout 0.1, Adam at 10^{-4}, batch 8, at most 50 epochs, early stopping on validation loss with patience 5). Three details follow the code rather than the paper: the reported F1 is Keras' `f1_m`, thresholded at 0.5 on each mini-batch of eight and averaged over batches; the classifier has a hidden layer rather than being linear; and oversampling also applies to the pre-training data. The missing-sensor evaluation of CroSSL does not run as released, so we implemented it as the paper describes: "missing at test" averages over all 15 subsets obtained by removing zero to three sensors, and "missing in training" removes a random subset from each training batch.

SleepSMC states that it follows the DeepSleepNet preparation on all 78 subjects; the pipeline above, applied to all 153 nights, gives 195,479 epochs, the count SleepSMC reports.[1] Sensors are the two EEG channels, the EOG and the chin EMG. SleepSMC releases neither code nor folds, so we draw its five subject-wise folds with a fixed seed (16, 16, 16, 15 and 15 test subjects) and hold out 20% of the remaining subjects for validation. The model is trained on all sensors and tested on all of them and on each modality alone.

MultiConsSleepNet uses Fpz-Cz EEG and horizontal EOG from the first 20 subjects with one subject held out per fold. It reports 44,220 epochs without stating its cropping rule and does not release its split; we use the 42,308 epochs of the DeepSleepNet preparation above and, in each fold, hold out two of the remaining 19 subjects for validation.

We follow MMBind (Ouyang et al., 2025): 14 actions, the first 30 frames of each sequence, and its subject split. Unsupervised training uses the 224 sequences of S01, S02, S11, S12, S21, S22, S31, S32, S03, S04, S13, S14, S23, S24, S33 and S34 with keypoints, depth and radar; the classifier is trained on the 112 sequences of S05, S06, S15, S16, S25, S26, S35 and S36; accuracy is measured on the 224 sequences of the remaining 16 subjects from depth and radar and, as in MMBind, averaged over epochs 81 to 190 of classifier training. Depth is reduced to 48×64 by 10×10 area averaging, radar keeps up to 92 points per frame with the MMBind normalisation, and keypoints are the 17 two-dimensional joints of the public release. Keypoints are taken relative to the frame-0 mid-hip and divided by the torso length, and depth relative to the per-pixel temporal median of its own 30 frames; neither transform uses labels or dataset statistics. The contrastive baseline is the MMBind multimodal objective with 128-dimensional projections, trained with the same encoders, inputs, subject batches, optimiser and number of epochs, and read by the same fusion head over its 30 frame features per sensor. Our fine-tuned variant trains the encoders together with the fusion head over the map, with the prototype tables fixed, Adam at 3×10^{-4} for 200 epochs.

The dataset has 861 instances of 27 actions by 8 subjects, each recorded simultaneously by a wrist-worn IMU and a Kinect (Chen et al., 2015). The inertial signal is resampled to $(6, 50)$ without scaling or centring; the skeleton is taken relative to the frame-0 spine joint, divided by the frame-0 shoulder distance, and resampled to $(3, 50, 20)$. We use a stratified cross-instance split of $603/129/129$

[1] The DeepSleepNet script deletes one night entirely (SC4362F0) when a recording ends on an exact 30-second boundary; keeping the complete epochs of that night reproduces the published count.

instances over five paired random partitions. Stratification makes micro accuracy, macro-recall and macro-F1 agree to within 0.30 points; we report micro accuracy.

C HYPERPARAMETERS AND REPRODUCTION

Table 9: Unsupervised training settings per benchmark.

	UTD-MHAD	Sleep-EDF-78	Sleep-EDF-20	MM-Fi
Encoders	1D CNN, HCN	DeepSleepNet CNN	DeepSleepNet CNN	MMBind
Widths d_m	128, 256	256, 256	256	16, 64, 64
K	400	400	400	100
Map temperature τ	0.1	0.03	0.03	0.1
Readout temperature τ_r	0.1	0.1	0.1	0.1
Grid weight λ	1	30	50	2/3
Negatives	batch	same recording	same recording	same subject
Epochs, batch	1000, 64	100, 256	100, 256	300, 14
Head	MLP	DeepSleepNet seq.	CroSSL MLP	MMBind fusion

Table 9 lists the settings. Shared by all benchmarks: per-sensor tables, cosine BMU, ℓ_2-normalised embeddings, NT-Xent on activation maps, the neighbourhood gradient reaching the encoder, $\sigma : 3.0 \to 0.5$ linear, Adam with cosine decay from 10^{-3} to 10^{-5}, no augmentation, contrastive temperature τ_c=0.1, and tables initialised uniformly in $\pm d_m^{-1/2}$. MM-Fi instead uses the MMBind optimiser (Adam at 5×10^{-4} without decay, weight decay 10^{-4}) and τ_c=0.07. Under the protocols of SleepSMC and MultiConsSleepNet the settings are those of the Sleep-EDF columns, with λ=10 and λ=30 and the DeepSleepNet sequence head. On Sleep-EDF a batch holds 32 consecutive epochs from each of 8 recordings, which supplies the same-recording negatives. The sequence head is the residual block of DeepSleepNet (a 1024-unit shortcut and a two-layer bidirectional LSTM with 512 units per direction) over 25 consecutive epochs, trained with Adam at 10^{-4} and gradient clipping at 10; evaluation sums log-probabilities over overlapping windows. On MM-Fi unsupervised training uses bfloat16 mixed precision, and the head is trained for 200 epochs with Adam at 3×10^{-4}, weight decay 10^{-4} and batch 32. `proto_head` additionally trains the tables at 10^{-4} with the unsupervised objective retained at weight 0.01, a setting the UTD-MHAD results are insensitive to (four values spanning three orders of magnitude differ by 0.8 points). The UTD-MHAD head is an MLP $K \to 256 \to 27$ trained for 200 epochs with weight decay 10^{-4}; its unsupervised checkpoint is fixed at epoch 1000 and was never selected with labels.

D ADDITIONAL RESULTS

Sleep-EDF-78. Test accuracy per fold is 84.9, 87.7, 80.6, 79.7, 84.7, 84.0, 85.5, 90.0, 84.4 and 82.1 (standard deviation 3.1). Per-class F1 on the pooled confusion matrix:

Model	Test input	W	N1	N2	N3	REM
PESTO (`proto_head`)	EEG, EOG	93.0	48.3	85.7	77.1	86.0
PESTO, EOG removed	EEG	92.5	45.5	85.2	77.8	81.6
XSleepNet, trained on EEG (Phan et al., 2022)	EEG	93.3	49.9	86.0	78.7	81.8

Sleep-EDF maps without a classifier. Agreement between the BMUs of two sensors of the same test epoch, against the same sensors paired with a random other epoch, and direct grid readout (each cell labelled by the majority stage of its training epochs) through the consensus cell and through each sensor alone. Models of Figure 2.

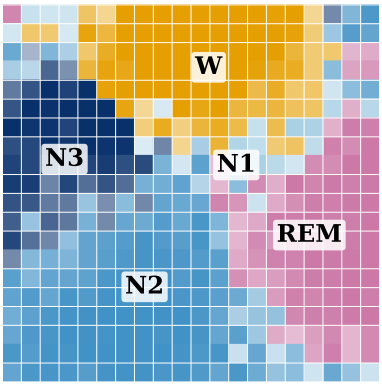


Figure 2: The unsupervised Sleep-EDF-20 map.
Two EEG channels, EOG and chin EMG, drawn as in Figure 1b: each of the 20 × 20 cells takes the majority sleep stage of the training samples it wins and is lighter where they are mixed, with letters at the stage centroids. This protocol has a single fixed split rather than folds; first seed.

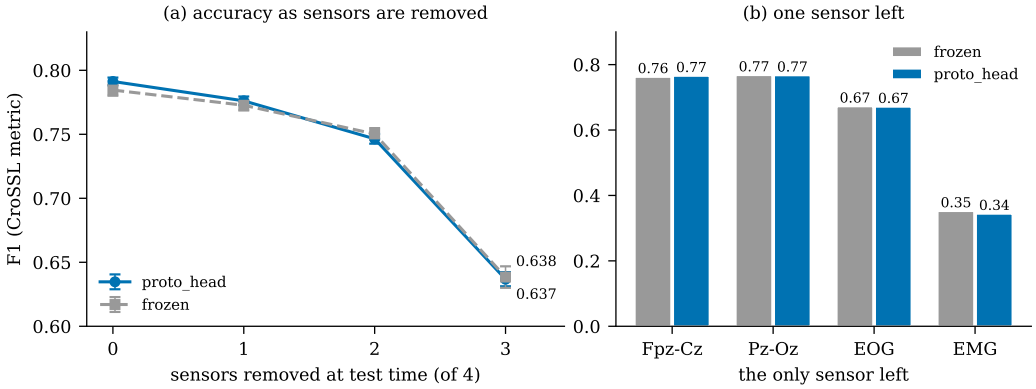


Figure 3: Removing sensors at test time, Sleep-EDF-20.
From a model trained on all four sensors, F1 as computed by CroSSL, mean over 5 seeds (error bars: s.d.). The map averages the affinities of the sensors that remain. **(a)** Averaged over all subsets with the given number of sensors removed. **(b)** Only one sensor left.

		same cell		within one cell	
Benchmark	Sensor pair	paired	unpaired	paired	unpaired
Sleep-EDF-78	EEG–EOG	0.178	0.002	0.332	0.024
Sleep-EDF-20	Fpz-Cz–Pz-Oz	0.236	0.002	0.372	0.024
	Fpz-Cz–EOG	0.133	0.003	0.272	0.026
	Pz-Oz–EOG	0.105	0.004	0.230	0.024
	EEG or EOG–EMG	0.011–0.012	0.002–0.003	0.046–0.059	0.019–0.022

Direct grid readout	consensus	Fpz-Cz	Pz-Oz	EOG	EMG	majority class
Sleep-EDF-78	0.757	0.748	—	0.730	—	0.257
Sleep-EDF-20	0.760	0.733	0.741	0.669	0.381	0.402

Sleep-EDF-20: other metrics. With all sensors, mean over 5 seeds with our encoder:

	CroSSL F1	macro-F1	accuracy	κ
PESTO, frozen	0.785	0.722	0.788	0.714
PESTO, `proto_head`	0.791	0.725	0.791	0.719

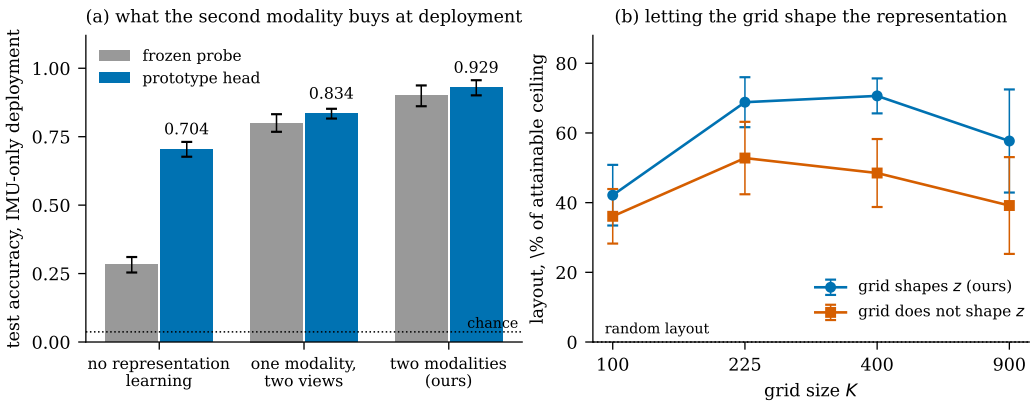


Figure 4: What the second modality and the map layout contribute, UTD-MHAD.
(a) Deployed on the inertial stream alone, varying only what was available during unsupervised training: representation learning and the second modality contribute separately. **(b)** Layout of the map as a percentage of the ceiling attainable in each run, with and without the neighbourhood gradient reaching the encoder. Error bars are standard deviations over five paired splits.

Table 10: Where the neighbourhood gradient is allowed to reach the encoder, UTD-MHAD.

K	layout %		Δ layout	splits	Δ `acc_strict`
	grid shapes z	it does not			
100	42.1	36.1	+6.1	4/5	+0.002
225	68.8	52.8	+16.0	**5/5**	−0.003
400	**70.6**	48.5	**+22.1**	**5/5**	−0.028
900	57.7	39.2	+18.5	**5/5**	−0.029

The gradient reaches the encoder or is stopped at the prototype tables, five paired splits, all other settings identical. Layout is a percentage of the ceiling attainable in each run; "splits" counts the paired comparisons that favour the first arm.

E GRID SIZE

Table 11: Grid size on Sleep-EDF-78, validation subjects of fold 0, `proto_head`, softmax readout. "EEG" removes the EOG at test time.

K	100	225	400	900	1600
EEG, EOG: Acc. / MF1	80.8 / 75.2	80.9 / 75.3	80.9 / 75.2	80.8 / 75.4	80.9 / 75.7
EEG: Acc. / MF1	80.9 / 75.5	80.6 / 76.1	81.0 / 75.6	81.0 / 76.8	81.1 / 76.3

On Sleep-EDF-78 (Table 11), accuracy is flat across the range: the five grid sizes span 0.2 accuracy points with both sensors and 0.4 with the EEG alone. On UTD-MHAD (Table 12), the grid read directly improves steadily with K, while the four `proto_head` values lie within the standard error of each other. Layout peaks in the middle of the range and falls at K=900, where only two thirds of the grid is occupied.

F READING THE MAP

The affinities of the present sensors are averaged before the softmax, so both readouts are defined over the same K cells of the map and handle missing sensors in the same way. On the validation subjects of Sleep-EDF-78 the four settings lie within 0.9 accuracy points and 2.0 macro-F1 points of one another (Table 13); we keep τ_r=0.1 on every benchmark, since that spread is smaller than the fold-to-fold spread of the protocol itself (a per-fold accuracy s.d. of 3.1 points, Appendix D). On the test subjects of Sleep-EDF-20 the two readouts reverse (Table 14): the softmax readout is ahead or level in every configuration, and the gap grows as sensors are removed. We therefore report the softmax readout throughout.

Table 12: Grid size on UTD-MHAD, test set, five paired splits.

K	acc_strict ↑	frozen ↑	proto_head ↑	layout % ↑
100	0.6155	0.9178	$0.9674 \pm .013$	48.1
225	0.7612	0.9612	$0.9783 \pm .013$	**72.7**
400	0.8744	**0.9659**	$\mathbf{0.9814 \pm .007}$	71.8
900	**0.9426**	0.9535	$0.9752 \pm .020$	58.7

Table 13: Readout of the map on Sleep-EDF-78, validation subjects of fold 0; only the input of the head changes. "Affinities" reads the averaged cosine affinities without a softmax.

Head input	EEG, EOG: Acc. / MF1 / κ	EEG: Acc. / MF1 / κ
softmax, τ_r=0.1 (reported)	80.9 / 75.2 / 0.745	81.0 / 75.6 / 0.746
softmax, τ_r=0.3	81.0 / 76.3 / 0.747	80.8 / 75.2 / 0.744
softmax, τ_r=1.0	80.3 / 74.3 / 0.737	80.2 / 75.7 / 0.737
affinities	81.2 / 75.5 / 0.749	80.9 / 76.2 / 0.746

G SINGLE-MODALITY BASELINES

A multimodal method should be measured against a single-modality self-supervised baseline on the same deployment; in the cross-subject results of Choi et al. (2023) on this dataset, three of four multimodal methods lose to single-modality SimCLR. Our baseline uses one encoder, one projection and one prototype table, run on two augmented views, with everything else identical to the multimodal arm. On the inertial deployment, learning a representation is worth 13.0 points over a random encoder and the second modality a further 9.5, with the same sign on every split (Table 15). The inertial augmentations are taken unmodified from Choi et al. (2023). Training on one modality with only the neighbourhood term, whose forces are all attractive, collapses completely: the mean pairwise cosine between prototypes reaches 1.0000 and the loss falls to exactly zero within 200 epochs. Collapse is the global optimum of an objective with no repulsive term; classical SOMs avoid it because their input is fixed. Restoring the in-batch negatives of the contrastive term restores prototype diversity (mean cosine 0.27–0.29). We therefore report the random-encoder arm as a floor rather than as a single-modality method.

H GRADIENT ROUTING

On the prototype table the alignment term is $60\times$ stronger than the neighbourhood term (Table 16), so grid structure has almost no say in how $\mathbf{P}$ is written. Removing the stop-gradient opens a second route for it, through the encoder.

Table 14: Softmax versus affinity readout on Sleep-EDF-20, test set, mean over 5 seeds. F1 as computed by CroSSL, by number of sensors removed.

		sensors removed				macro-F1	
Protocol	Readout	0	1	2	3	all	missing
frozen	softmax	**0.785**	**0.773**	**0.750**	**0.637**	**0.722**	**0.656**
	affinities	0.779	0.767	0.741	0.630	0.710	0.644
`proto_head`	softmax	**0.791**	**0.776**	**0.746**	**0.638**	**0.725**	**0.652**
	affinities	0.784	0.758	0.705	0.600	0.715	0.608

Table 15: UTD-MHAD, test set, five paired splits, deployed on the inertial stream. Only what was available during unsupervised training differs down the rows.

Unsupervised training signal	frozen	`proto_head`
Two modalities (inertial + skeletal)	$\mathbf{0.8992}$	$\mathbf{0.9287}$
Inertial only, two augmented views	0.8000	0.8341
No representation learning (floor)	0.2822	0.7039
Two modalities − two views	+0.0992	+0.0946

Table 16: Gradient norm reaching each parameter group when each term is back-propagated alone, UTD-MHAD, trained model, one batch of 64.

Term	$\rightarrow$ encoder	$\rightarrow$ projection	$\rightarrow \mathbf{P}$
① cross-modal alignment	2.7733	1.4972	1.3027
③ Gaussian neighbourhood	0.1307	0.3079	0.0218