InfoMAE: Pair-Efficient Cross-Modal Alignment for Multimodal Time-Series Sensing Signals

Tomoyoshi Kimura
tkimura4@illinois.edu
University of Illinois Urbana-Champaign
Urbana, IL, USA

Xinlin Li
Osama Hanna
xinlinli@g.ucla.edu
ohanna@ucla.edu
University of California, Los Angeles
Los Angeles, CA, USA

Yatong Chen
chenyatong@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Yizhuo Chen
Denizhan Kara
yizhuoc@illinois.edu
kara4@illinois.edu
University of Illinois Urbana-Champaign
Urbana, IL, USA

Tianshi Wang
Jinyang Li
tianshi3@illinois.edu
jinyang7@illinois.edu
University of Illinois Urbana-Champaign
Urbana, USA

Xiaomin Ouyang
xmouyang@cse.ust.hk
Hong Kong University of Science and Technology
Hong Kong SAR, China

Shengzhong Liu
shengzhong@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Mani Srivastava
Suhas Diggavi
mbs@ee.ucla.edu
suhas@ee.ucla.edu
University of California, Los Angeles
Los Angeles, CA, USA

Tarek Abdelzaher
zaher@illinois.edu
University of Illinois Urbana-Champaign
Urbana, IL, USA

Abstract

Standard multimodal self-supervised learning (SSL) algorithms regard cross-modal synchronization as implicit supervisory labels during pretraining, thus posing high requirements on the scale and quality of multimodal samples. These constraints significantly limit the performance of sensing intelligence in IoT applications, as the heterogeneity and the non-interpretability of time-series signals result in abundant unimodal data but scarce high-quality multimodal pairs. This paper proposes InfoMAE, a cross-modal alignment framework that tackles the challenge of multimodal pair efficiency under the SSL setting by facilitating efficient cross-modal alignment of pretrained unimodal representations. InfoMAE achieves *efficient cross-modal alignment* with *limited data pairs* through a novel information theory-inspired formulation that simultaneously addresses distribution-level and instance-level alignment. Extensive experiments on two real-world IoT applications are performed to evaluate InfoMAE's pairing efficiency to bridge pretrained unimodal models into a cohesive joint multimodal model. InfoMAE enhances downstream multimodal tasks by over 60% with significantly improved multimodal pairing efficiency. It also improves unimodal task accuracy by an average of 22% [1].

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Learning paradigms**.

Keywords

Multimodal sensing, Self-supervised learning, Internet of Things

ACM Reference Format:
Tomoyoshi Kimura, Xinlin Li, Osama Hanna, Yatong Chen, Yizhuo Chen, Denizhan Kara, Tianshi Wang, Jinyang Li, Xiaomin Ouyang, Shengzhong Liu, Mani Srivastava, Suhas Diggavi, and Tarek Abdelzaher. 2025. InfoMAE: Pair-Efficient Cross-Modal Alignment for Multimodal Time-Series Sensing Signals. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3696410.3714853

1 Introduction

Multimodal Self-Supervised Learning (SSL) algorithms, although achieving unprecedented performance in extensive sensing applications [11, 12, 32, 52], present unique data challenges rarely encountered with unimodal SSL or vision-language domains due to the complexity in acquiring high-quality multimodal pairs for IoT signals. The inherent properties of sensory data common in Web and Industrial sensing applications result in abundant unimodal signals but scarce multimodal pairs. First, sensory modalities have

[1]The code is available at https://github.com/tomoyoshki/InfoMAE.

WWW '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714853

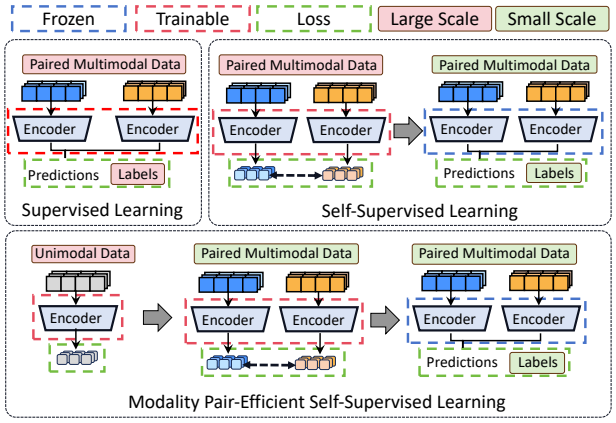

Figure 1: Comparison of supervised learning, self-supervised learning, and pair-efficient self-supervised learning.

heterogeneous properties, such as sampling rate, timestamp, or duration, that increase the likelihood of capturing asynchronous events. For example, in vibration sensing applications (machine monitoring, vehicle detection), multimodal sensors (geophone, microphone, thermometer, etc.) often operate at different sampling rates, leading to temporal misalignments that require manual calibration [40]. Second, raw signals often lack intuitive interpretability. Unlike images or text, where visual features can be easily matched to textual captions, capturing useful signatures between sensing modalities like motion or frequency waves is challenging. Preprocessing and calibrating these signals requires modality-specific domain knowledge, which is labor-intensive and susceptible to operational errors. Finally, sensors for IoT are subject to varying deployment conditions, leading to sparse and noisy data [36]. For example, in human activity recognition (HAR) applications, wearable IMU sensors generate multimodal motion streams for real-time monitoring, fitness tracking, or healthcare purposes. Each modality can be independently affected by device constraints, platform heterogeneity, sensor failures, or variations in deployment environments, leading to missing or incomplete data streams. This heterogeneity often yields poor-quality uncorrelated multimodal pairs or incomplete datasets with significant gaps and missing data. As IoT networks scale in quantity and the number of modalities, acquiring large-scale, high-quality multimodal pairs becomes increasingly time-consuming, error-prone, and less reliable.

Despite these challenges, most existing multimodal SSL frameworks [1, 35, 48, 55] rely heavily on massive multimodal pairs to learn robust joint representations during the pretraining, but their capability could degrade significantly with insufficient synchronized pairs [44, 71] or uninformative false-positive pairs [9, 49] On the other hand, independently pretraining each modality on their unimodal data and directly concatenating misaligned modality features for finetuning fails to capture cross-modal interactions that are critical to downstream multimodal tasks [27, 68]. Instead, we observe that with limited multimodal pairs, we can effectively convert independently trained unimodal encoders into a coherent model that sustains strong generalizability in multimodal tasks. We refer to this process as *pair-efficient SSL*. The relation of pair-efficient SSL for multimodal data compared to standard SSL draws an analogy to the evolvement of SSL compared to supervised learning, as visualized in Figure 1. In supervised learning, manual labels serve as supervision to train encoders for mapping inputs to task-specific labels. Its performance depends heavily on the quantity and quality of human annotations. Self-supervised learning (SSL) mitigates label scarcity by first designating proxy labels from the data properties to learn general semantics with massive unlabeled data, then calibrating the pretrained model to a downstream task with minimal human annotations. Similarly, in multimodal SSL contexts, cross-modal alignment acts as a special form of "supervision", where point-to-point modality correspondence is utilized to identify semantically meaningful and consistent sensory information. Taking another step forward, pair-efficient SSL takes advantage of abundant unimodal data for "independent pretraining", followed by "cross-modal finetuning" with limited multimodal pairs to align unimodal models into a cohesive multimodal model.

In this paper, we propose InfoMAE, a cross-modal learning framework designed to enhance the alignment of unimodal representations using a limited number of multimodal pairs. The key idea behind InfoMAE is to enforce alignment across modalities at both the *distribution* and *instance* levels. Existing contrastive learning frameworks adopt point-to-point alignment to map samples across different modalities to a proximate joint representation [40, 51, 54, 62]. These approaches focus on aligning individual samples, essentially viewing alignment as a local optimization problem that aims to minimize the geometric distances between corresponding samples in the representation space. However, such instance-level approaches face significant challenges with limited multimodal pairs, as they may overfit to the specific pairs available and result in poor generalization with pairing biases. These hinder capturing complex cross-modal relationships, especially when the multimodal pairs are sparse and unevenly distributed. In contrast, InfoMAE takes a more holistic approach by emphasizing *distribution-level* alignment, considering the overall information content of the limited multimodal pairs rather than only focusing on the individual samples. We present a comprehensive analysis of distribution alignment and propose an *information theory-based approach* to formally define the distribution alignment problem in the factorized information space. We formulate this as a differential learning objective to construct (i) shared joint representations as a compact common variable across modalities capable of performing any multimodal task and (ii) private representations holding implicit modality-specific information independent of shared representations. InfoMAE alleviates the strict requirement of exact multimodal sample pairs and can better accommodate potential misalignments in data collection or temporal synchronization, improving the representations learned even with a small-scale multimodal pair.

We extensively evaluate InfoMAE across various combinations of pretrained unimodal domains. InfoMAE achieves exceptional performance gain compared to the standard multimodal SSL paradigm under limited multimodal pairs and outperforms existing works when aligning the unimodal representations. Individual unimodal encoders, in return, can also benefit from the representational structures with improved downstream performance. Additionally, as the number of multimodal pairs scale, InfoMAE also demonstrates versatility as a standard multimodal SSL framework, achieving SOTA performance across real-world IoT applications.

2 Analysis of Cross-Modal Alignment

2.1 Notation

Consider M sets of unsynchronized sensory modality data $\mathcal{X} = \{X_i\}_{i\in M}$, where each set X_i contains unlabeled samples of fixed-length windows partitioned from the time-series signals of the i-th sensory modality. Let $N_i = |X_i|$ denote the size of each set.

For the j-th sample of modality set i, we apply Short-Time Fourier Transform (STFT) to obtain its time-frequency representation, $\mathbf{x}_{ij} \in \mathbb{R}^{C_i \times I \times S_i}$, where C_i is the number of input channels, I is the number of time intervals within a sample window, and S_i is the spectrum length in the frequency domain. We have a set of modality encoders $\mathcal{E} = \{E_1, E_2, \ldots, E_M\}$ to extract the modality embeddings of each sample and a set of modality decoders $\mathcal{D} = \{D_1, D_2, \ldots, D_M\}$ to map the samples from the embedding space back to the time-frequency domain $\hat{\mathcal{X}} = \{\hat{X}_i\}_{i\in M}$ as a part of the reconstruction process. Additionally, there is a set of multimodal data $\mathcal{X}^s = \{X_i^s\}_{i\in M^s}$ consisting of a subset of modalities $M^s \subseteq \hat{M}$, where samples across the modalities are synchronized in time and have equal sizes $|X_1^s| = \cdots = |X_{M^s}^s|$. Note that each synchronized data of modality i can also be a subset of the unsynchronized unimodal set such that $X_i^s \subseteq X_i$, as any synchronized multimodal data is inherently unsynchronized when considered independently. Finally, we have a set of labeled data for supervised learning and finetuning on a much smaller scale, where each sample has a corresponding label y_j for each downstream task.

2.2 Problem Definition

Prior multimodal SSL practices rely on large-scale, fully synchronized multimodal sets $\mathcal{X}^s$ to learn joint multimodal representations for downstream tasks. However, these approaches overlook two challenges: (i) *Insufficient multimodal data*: When $|\mathcal{X}^s|$ is small, existing methods struggle to learn effective joint representations, and (ii) *Unutilized unimodal data*: The abundance of unimodal data is often ignored. In IoT applications, synchronized multimodal sets are limited due to signal heterogeneities, temporal misalignment, or domain variances, leading to incomplete modalities. This results in limited synchronized multimodal data compared to unimodal data ($|X_i^s| \leq |X_i|$). To better leverage unimodal data, our problem falls under the SSL setting with unimodal pretrained models and limited multimodal pairs, consisting of two stages:

Stage 1: Independent Unimodal Pretraining. For each independent modality data X_i, we train a corresponding unimodal encoder E_i. The goal is to learn a *holistic unimodal representation* that maximizes downstream unimodal performance after finetuning. Since modality sets X_i are independent, this pretraining is not limited by the number of synchronized pairs and can, therefore, fully leverage the abundant unimodal data.

Stage 2: Efficient Cross-Modal Alignment. Given a set of synchronized modalities data $\mathcal{X}^s$ of $M^s \subseteq M$ modalities, we aim to align the pretrained encoders efficiently. This alignment projects unimodal representations into joint representations that maximize the downstream multimodal performance after finetuning. The scale of the multimodal alignment should be significantly smaller than the unimodal pretraining $|X_i^s| \lll |X_i|$. In contrast to prior multimodal SSL works focusing on learning robust joint representations on large-scale multimodal data, this work aims to improve the *data efficiency* of learning robust joint representations given only limited multimodal pairs.

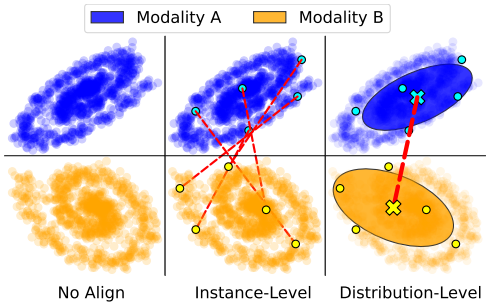

Figure 2: An illustration of instance-level vs. distribution-level Cross-Modal Alignment

2.3 Factorization & Distributional Alignment

This section analyzes multimodal representation factorization in the information space and demonstrates how it enables distribution-level alignment of unimodal representations.

2.3.1 Connection between Factorization and Cross-modal Alignment. In aligning multimodal representations, prior approaches often rely on contrastive learning to minimize the *modality gap* [39] by pulling representations of different modalities from the same sample closer together while pushing representations from different samples further apart. However, due to the inherent heterogeneity, each modality contains unique, modality-specific information, and enforcing perfect alignment across modalities could potentially hurt the performance in multimodal downstream tasks [28]. To address these challenges, recent works [28, 37, 40] have proposed factorizing modality representations into shared and private subspaces. It preserves both common and modality-specific information and allows for the alignment of shared representations while maintaining independent private representations for downstream tasks. However, these works operate on *instance-level alignment* and do not explore scenarios with limited multimodal data. The scarcity of paired samples introduces the risk of biased sampling, potentially misleading the alignment process. With this in mind, we analyze a different approach that factorizes the representation in the information space and enforces *distribution-level* alignment to capture a more comprehensive correlation between modalities by *emphasizing their information content rather than just their geometric proximity*. The intuition behind this is that instead of individual sample pairs, we aim to align modalities by the global structure (as shown in Figure 2). When the multimodal pairs are scarce, the distributional alignment aims to be *resilient to sampling biases* and capture meaningful cross-modal relationships.

2.3.2 Distributional Alignment through Information-theory based Factorization. We now formally define the factorization problem in the information space. Without loss of generality, we state the definitions for two modalities, $\mathcal{X} = \{X_1, X_2\}$, but they can be generalized to more modalities.

First, we are interested in constructing a compact random variable U (shared representation) that can perform any task that can

be achieved using X_1 separately and X_2 separately. Formally, we define a sufficient common variable as follows.

Definition 2.1. (Sufficient Common Variable) U is defined as the sufficient common variable between X_1, X_2 if and only if $U = g_1(X_1) = g_2(X_2)$ for some g_1, g_2, and

$$(\forall f_1, f_2)\Big([f_1(X_1) = f_2(X_2)] \implies [(\exists f) f(U) = f_1(X_1) = f_2(X_2)]\Big), \tag{1}$$

namely, any common (shared) function between X_1, X_2 can be computed using U. Building on the sufficient common variable, we define the shared representation to be the most compact form of U with the minimized entropy to ensure that U captures only the essential shared features across modalities.

Definition 2.2. (Shared Representation) We refer to a sufficient common variable U with minimal entropy $H(U)$ as the shared representation.

However, it is not clear how to find a sufficient common variable or a shared representation. We show that an approximation of the shared representation can be obtained by solving the following optimization problem, and later in Section 3, we propose the differentiable loss objectives with proof provided in Appendix A.

$$\min H(U) \text{ s.t. } X_1 \perp\!\!\!\perp X_2 \mid U, (\exists s_1, s_2)\ U = s_1(X_1) = s_2(X_2) \tag{2}$$

The conditional independence in Equation 2 enforces a form of distributional alignment, ensuring that given the shared representation U is the most compact aligned representation such that X_1, X_2 provide no additional information about each other. Moreover, we define the private representations V_1, V_2 between X_1, X_2 as follows.

Definition 2.3. (Private Representation) V_1, V_2 is the private representation of X_1, X_2 if they have minimal entropy among the random variables satisfying: $V_1 = p_1(X_1), V_2 = p_2(X_2)$ for some p_1, p_2 and there exist functions g_1, g_2 such that $X_1 = g_1(V_1, U), X_2 = g_2(V_2, U)$, where U is the shared representation.

Similarly, we look for approximate representations. In particular, we replace equalities with a distance constraint d, and independence is replaced by small mutual information. In Section 3, we discuss the detailed implementation of a differentiable loss function to find the approximate representations.

3 InfoMAE

This section introduces InfoMAE, a novel cross-modal alignment framework that efficiently aligns unimodal representations at the distribution and instance levels. We provide a detailed overview of InfoMAE's cross-modal alignment module in Figure 3.

3.1 Unimodal Pretraining

Unlike standard multimodal SSL that pretrains on synchronized multimodal pairs, we first initiate *unimodal pretraining* on large-scale unsynchronized unimodal data. In the first stage, we pretrain each encoder E_i independently on unimodal data X_i with masked reconstruction, defined as the following for each modality $i \in M$:

$$\mathcal{L}_i^{\text{unimodal}} = ||\hat{X}_i - X_i||^2 \mid \hat{X}_i = D_i(E_i(X_i)). \tag{3}$$

The pretrained unimodal encoders E_i extract a generalized representation for each modality M_i. However, they do not guarantee information compatibility between modalities when used together in the downstream tasks. In the following sections, we present InfoMAE's different components (as illustrated in Figure 4) to calibrate the encoders to *explicitly align* the modalities in both the distribution-level and instance-level with only a limited amount of multimodal pair $\mathcal{X}^s$.

3.2 Distribution-level Alignment

We begin with the differentiable objective function that we optimize to obtain the (approximate) shared (U) and private representations (V) defined in Section 2.3.2. To extract U that is a function of both X_1, X_2, we equivalently extract $U_1 = F_1^{\text{shared}}(E_1(X_1)), U_2 = F_2^{\text{shared}}(E_2(X_2))$, where F_1, F_2 are 2-layer MLP projectors that maps the general representation into factorized representations, and enforce a constraint that $U_1 = U_2$. Similarly, we extract $V_1 = F_1^{\text{private}}(E_1(X_1)), V_2 = F_2^{\text{private}}(E_2(X_2))$. $\mathcal{U} = \{U_1, U_2\}$ and $\mathcal{V} = \{V_1, V_2\}$ denote the shared and private representations, respectively.

3.2.1 Shared Representation. As described in Section 2, we aim to find the shared representation U that solves the optimization problem in Definition (2.2). However, due to the difficulty of the optimization problem [2] and the possibility that a shared representation does not exist, we instead approximate the shared representation by minimizing the following objective

$$\begin{aligned}\mathcal{L}_{\text{info}}^{\text{shared}} =& \alpha d(U_1, U_2) + \beta(H(U_1) + H(U_2)) \\ &+ I(X_1; X_2 \mid U_1) + I(X_1; X_2 \mid U_2),\end{aligned} \tag{4}$$

where α and β are the hyperparameters controlling the weight of each term, and $d(\cdot)$ is a distance measure. The first two terms in the loss function aim to find $U_1 = U_2$ with minimal entropy, while the last two terms aim to impose conditional independence of X_1, X_2 given U_1 or U_2. We would like to note that the entropy and conditional mutual information listed in Eq. (4) are not easy to compute or differentiate. To alleviate this, we reduce these terms into probabilistic density functions below:

$$\begin{aligned}\mathcal{L}_{\text{info}}^{\text{shared}} = \alpha d(U_1, U_2) + \sum_{i=1}^{2} \mathbb{E}_{X_1, X_2, U_i} \Bigg[&\log \frac{p_{X_1, X_2, U_i}}{p_{X_1} p_{X_2} p_{U_i}} \\ &+ (1-\beta) \log \frac{p_{X_i, U_i}}{p_{X_i} p_{U_i}} + \log \frac{p_{X_{3-i}, U_i}}{p_{X_{3-i}} p_{U_i}} \Bigg].\end{aligned} \tag{5}$$

Due to the space limit, we leave the detailed proof and discussion in Appendix A. To further enhance the differentiability of Eq. (5) by avoiding directly computing the probabilistic density (*e.g.*, $\log \frac{p_{X_1, X_2, U_i}}{p_{X_1} p_{X_2} p_{U_i}}$), we follow [31, 50, 59] and utilize the *density-ratio trick* to train a discriminator $\mathcal{R}$, which given X_1, X_2, U, outputs the probability that X_1, X_2, U are generated from p_{X_1, X_2, U_i}, instead of $p_{X_1} p_{X_2} p_{U_i}$. The density ratio can then be estimated as

$$\log \frac{p_{X_1, X_2, U_1}}{p_{X_1} p_{X_2} p_{U_1}} = \log \frac{\mathcal{R}(X_1; X_2; U_1)}{1 - \mathcal{R}(X_1; X_2; U_1)}. \tag{6}$$

We train the discriminators jointly with the encoders and describe the training for both in Appendix D.

[2] The optimization problem in Definition (2.2) is non-convex with a possibly infinite number of variables.

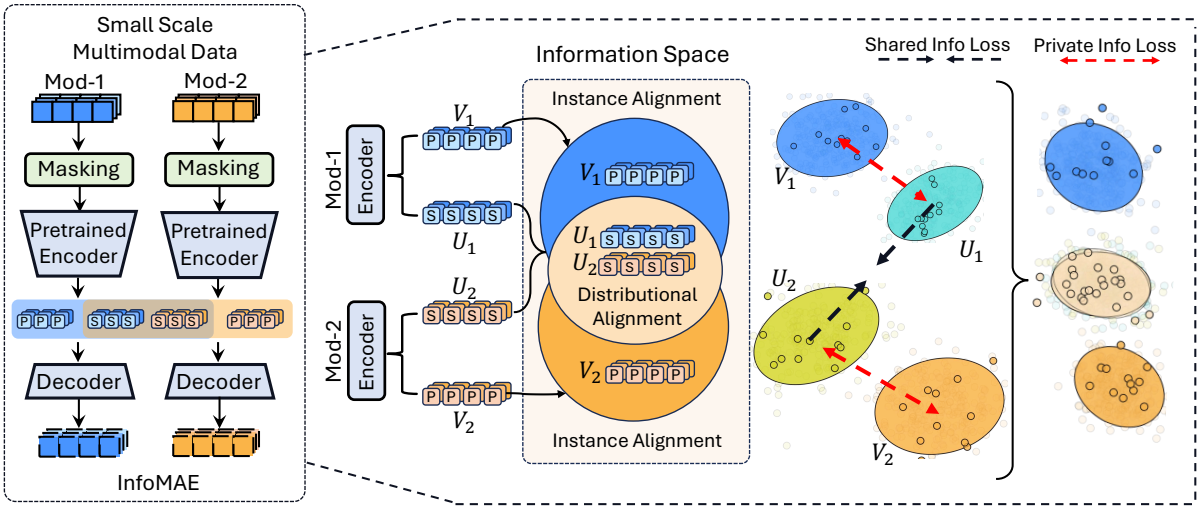

Figure 3: Overview of InfoMAE's alignment in the information space. InfoMAE adopts an information theory-inspired objective to align the factorized representations. Best viewed in color.

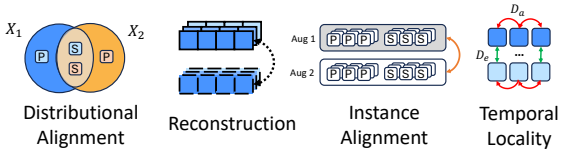

Figure 4: Key learning objectives of InfoMAE.

3.2.2 Private Representation. As the decoders take both the shared and private representations as input, the self-reconstruction objective would enforce the private representations V to capture the implicit modality-specific information. Following Definition 2.3, we minimize the entropy of the private representations (V_1, V_2). In addition, for each modality, we expect the private and shared representations to be independent. To better guide the learning process, we explicitly minimize their mutual information. The objectives of the private representations can be summarized as the following:

$$\mathcal{L}_{\text{info}}^{\text{private}} = \gamma H(V_1) + \gamma H(V_2) + \epsilon I(V_1; U_1) + \epsilon I(V_2; U_2), \quad (7)$$

where γ and ϵ are used as the hyperparameters for private entropy and shared private independence. Similar to Eq.(5), we apply *density-ratio trick* (Eq.(6)) to estimate each term in Eq. (7).

While the formulation effectively aligns modality representations within the information space, it depends on further learning objectives to ensure they are meaningful for downstream tasks. Next, we will describe the additional components of InfoMAE that are designed to capture meaningful representations.

3.3 Self Reconstruction

InfoMAE applies the masked reconstruction objective to enforce that the learned representation captures the critical semantical information through reconstruction loss. Following MAE[23], we mask out 75% of the patched input. To ensure both the shared and private representation are meaningful, the decoder takes in the concatenated shared and private representations $\mathbf{h}_{ij} = \mathbf{u}_{ij} || \mathbf{v}_{ij}$ to reconstruct the input $\hat{\mathbf{x}}_{ij}$. We compute the MSE on the masked portion of the reconstructed $\hat{\mathbf{x}}_{ij}$ and the original input $\mathbf{x}_{ij}$ with δ as the hyperparameter and $D_i(\cdot)$ as the decoder for modality i.

$$\mathcal{L}_{\text{reconstruction}} = \delta \sum_{i \in M} \sum_{j \in B} ||\mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij}||^2 \mid \hat{\mathbf{x}}_{ij} = D_i(\mathbf{h}_{ij}). \quad (8)$$

3.4 Instance-level Alignment

Augmentations are primarily used to generate different views for private-space contrastive learning in most existing works [28, 37, 40]. However, we argue that the transformation invariance property should be reflected in both private and shared representations to understand the instance variances. Thus, InfoMAE adds a contrastive loss on the concatenated representation of the shared and private spaces $\mathbf{h}_{ij}$ by treating two randomly different augmented views as the positive pairs with λ and τ as the hyperparameters.

$$\mathcal{L}_{\text{aug}} = \lambda \sum_{i \in M} \sum_{j \in B} \log \frac{\exp\left(\mathbf{h}_{ij} \cdot \mathbf{h'}_{ij}/\tau\right)}{\sum_{k \neq j \in B} \exp\left(\frac{\mathbf{h}_{ij} \cdot \mathbf{h}_{ik}}{\tau}\right) + \sum_{k \in B} \exp\left(\frac{\mathbf{h}_{ij} \cdot \mathbf{h'}_{ik}}{\tau}\right)}. \quad (9)$$

3.5 Temporal Locality

We apply a simple ranking constraint to learn *temporal locality* of time-series signals. During pretraining, a sequence sampler randomly selects a batch of sequences consisting of a fixed number of consecutive samples, while the samples across sequences are distant in time. We define $C_{xy'} = \sum_{i=1}^{L} \sum_{j=1}^{L} d(x_i, y_j)$, as the average Euclidean distance (d) of all sample embedding pairs between the sequence s and s' of length L. Then, the temporal constraint with a hyperparameter η can be defined as::

$$\mathcal{L}_{\text{temp}} = \eta \sum_{s \in B} \sum_{s' \neq s \in B} \max\left(C_{ss} - C_{ss'} + 1, 0\right) \quad (10)$$

where C_{ss} and $C_{ss'}$ measure the average intra-sequence (D_a) and inter-sequence (D_e) distances . The added 1 is the margin indicating the minimum gap between the two distances. η is used as the hyperparameter to control the weight of the temporal constraint.

Finally, the overall training objective of InfoMAE for the cross-modal alignment stage can be summarized as follows:

$$\mathcal{L} = \mathcal{L}_{\text{info}}^{\text{shared}} + \mathcal{L}_{\text{info}}^{\text{private}} + \mathcal{L}_{\text{reconstruction}} + \mathcal{L}_{\text{aug}} + \mathcal{L}_{\text{temp}}. \quad (11)$$

Table 1: Linear probing performance of Moving Object Detection on domain M. We align pretrained unimodal encoders from different domains. $A_{Sei}||B_{Aco}$ means seismic encoder from domain A and acoustic encoder from domain B are aligned.

Framework	Aligned Domains		T_{Sei} || M_{Aco}		G_{Sei} || T_{Aco}		T_{Sei} || T_{Aco}		G_{Sei} || M_{Aco}		T_{Sei} || G_{Aco}	
	Joint Pretrain	Modal Alignment	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Unimodal Concat	✗	✗	0.6731	0.6699	0.5392	0.5281	0.4454	0.4366	0.7247	0.7217	0.6584	0.6543
CMC [62]	✗	✓	0.6792	0.6702	0.4313	0.4356	0.4173	0.4032	0.6919	0.6877	0.6497	0.6335
FOCAL [40]	✗	✓	0.7462	0.7432	0.6249	0.6249	0.5613	0.5579	0.7549	0.7527	0.7194	0.7160
GMC [54]	✗	✓	0.7354	0.7317	0.6591	0.6523	0.4756	0.4720	0.8044	0.8053	0.7247	0.7211
SimCLR [6]	✗	✓	0.3061	0.2742	0.2873	0.2609	0.2974	0.2758	0.2981	0.2698	0.2800	0.2308
TNC [63]	✗	✓	0.1969	0.0815	0.1788	0.1312	0.1855	0.1021	0.1929	0.0896	0.1949	0.1041
TSTCC [14]	✗	✓	0.3001	0.2706	0.2639	0.2393	0.2867	0.2432	0.3048	0.2842	0.2860	0.2337
InfoMAE	✗	✓	**0.7950**	**0.7929**	**0.6986**	**0.7007**	**0.5928**	**0.5908**	**0.8326**	**0.8324**	**0.7636**	**0.7537**
Joint Pretrain	✓	✗	Acc: 0.3329				F1: 0.3039					

InfoMAE adopts both distribution-level and instance-level alignment of each modality's factorized shared and private representations. Since the cross-modal alignment of InfoMAE is also a generalized multimodal framework, we would also like to note that this objective can be used as the joint multimodal pretraining objective.

4 Evaluation

4.1 Experimental Setup

4.1.1 Backbone Encoder. We adopt the SWIN Transformer (SW-T) [41] as the backbone encoder for our framework. SW-T computes local attention within shifted windows on input spectrogram patches to extract comprehensive time-frequency representations.

4.1.2 Datasets. Our experiments focus on Moving Object Detection (MOD) and Human Activity Recognition (HAR). The MOD application contains vibration-based datasets using seismic and acoustic sensors. The HAR application consists of publicly released IMU sensor datasets collected from human subjects performing various daily activities. To evaluate cross-modal alignment, we simulate a practical scenario where the pretrained domains differ significantly to reflect the diverse signals across different IoT domains. Under this setting, we have unsynchronized unimodal data from different domains: MOD consists of data from three separately collected domains (M, G, T), each with different targets, terrains, and environmental conditions. HAR consists of two datasets (RW-HAR [61] and PAMAP2 [56]). We pretrain unimodal encoders with only the unimodal data from each domain and then use small-scale synchronized multimodal pairs for cross-modal alignment. For joint pretraining, we pretrain on the massive available synchronized multimodal pairs. We summarize and describe these applications and domains in more detail in Appendix B.

4.1.3 Baselines. We compare InfoMAE with different SOTA SSL baselines including unimodal CL (SimCLR[7], MoCo[8]), multimodal CL (CMC[62], GMC[54], FOCAL [40]), temporal CL (TNC[63], TSTCC[14]), and MAE based frameworks (MAE[23], CAV-MAE[17]).

4.2 Cross-Modal Alignment Evaluation

4.2.1 Moving Object Detection. We evaluate InfoMAE against prior CL works [7, 14, 40, 54, 62, 63] on cross-modal alignment with various combinations of unimodal encoders (seismic and acoustic) pretrained with different domains. We align the encoders with a small scale of multimodal pairs (5% of the unimodal data scale) and an even smaller subset of labeled multimodal pairs from domain M for finetuning. MOD application involves two modalities (seismic and acoustic). Therefore we represent the domains of the unimodal representations with two letters (*e.g.*, $T_{\text{Sei}}||G_{\text{Aco}}$ represents aligning the seismic encoder pretrained on domain T and acoustic encoder pretrained on domain G).

In addition to the prior CL baselines, we also show the performance for direct concatenation of the pretrained unimodal representations without any alignment and for Joint Multimodal Pretraining on the same amount of synchronized multimodal pairs. We present the finetune accuracy and F1-score in Table 1, InfoMAE consistently outperforms the unimodal concatenation by a significant margin since direct concatenation fails to exploit cross-modal correspondence. CMC and other unimodal SSL frameworks even have negative impacts compared to direct concatenation, indicating that unimodal objectives or simply aligning the multimodal representations without considering the modality discrepancy could hurt the downstream performance. InfoMAE also achieves better results than FOCAL and GMC, underscoring the benefits of enforcing distribution-level alignment over instance-level alignment in downstream tasks with limited multimodal data. When the same amount of multimodal data is used for Joint Multimodal Pretraining, the significant gap between the aligned unimodal models and the joint pretrained multimodal model suggests the feasibility of transferring pretrained unimodal representations to multimodal representations with only limited (5%) synchronized multimodal data. Note that some domain combinations (*e.g.*, , $G_{\text{sei}}||T_{\text{aco}}$, $T_{\text{sei}}||T_{\text{aco}}$, $T_{\text{sei}}||G_{\text{aco}}$) do not even overlap with the alignment and finetuning domain M.

4.2.2 Human Activity Recognition. Besides MOD application, we also evaluate InfoMAE on HAR applications. In contrast to MOD evaluation, which aligns unimodal encoders pretrained on different domains, we analyze how additional unsynchronized data from the same domains could assist the downstream performance given the limited number of multimodal pairs. Here, we independently pretrain all unimodal encoders on unsynchronized IMU data from either PAMAP2, RW-HAR, or Combined, which is the concatenation of the former two. Then, we use a small portion of the synchronized multimodal data pairs from PAMAP2 for cross-modal alignment

Table 2: Alignment performance (MM) with different multimodal pair ratios from MOD.

Multimodal Data	Supervised		Joint Pretrain		CMC		GMC		FOCAL		**InfoMAE**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
5%	0.5740	0.5663	0.3329	0.3039	0.7087	0.6989	0.8614	0.8616	0.8694	0.8668	**0.8828**	**0.8808**
15%			0.6142	0.6104	0.8111	0.8062	0.8781	0.8753	0.8727	0.8703	**0.9049**	**0.9028**
25%			0.7071	0.7938	0.8433	0.8372	0.8774	0.8759	0.8848	0.8831	**0.9290**	**0.9270**
50%			0.8942	0.8920	0.8754	0.8724	0.8948	0.8938	0.9009	0.8994	**0.9377**	**0.9367**

Table 3: Linear probing performance of HAR on PAMAP2 by aligning pretrained unimodal encoders.

Unimodal Pretrain Domain	Combined		PAMAP2		RW-HAR	
Multimodal Alignment Domain	PAMAP2		PAMAP2		PAMAP2	
Metric	Acc	F1	Acc	F1	Acc	F1
Concat	0.7843	0.7000	0.7763	0.6210	0.5675	0.4187
CMC	0.7334	0.6508	0.7285	0.6788	0.7010	0.5956
FOCAL	0.7922	0.7129	0.7354	0.6327	0.7643	0.6243
GMC	0.7314	0.5915	0.7344	0.5869	0.7414	0.5816
SimCLR	0.7299	0.6190	0.7075	0.5426	0.7225	0.5581
TNC	0.5431	0.4080	0.5889	0.4824	0.6378	0.5167
TSTCC	0.7299	0.6003	0.7065	0.5773	0.7354	0.5864
InfoMAE	**0.8261**	**0.7303**	**0.8117**	**0.7175**	**0.7912**	**0.6901**

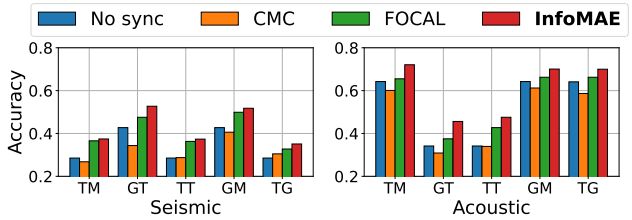

Figure 5: Unimodal linear probing accuracy of MOD with and without cross-modal alignment.

and downstream finetuning. We present the results in Table 3. InfoMAE consistently achieves the best performance, with an average of 4.09% and 5.16% improvements in accuracy and the F1-score compared to the best-performing baseline, FOCAL. The improvement is most significant in aligning unimodal encoders pretrained on RW-HAR, which completely differs from the alignment set (PAMAP2). This further demonstrates InfoMAE's robustness as an alignment framework with a limited amount of multimodal pairs, reflecting its superior ability to utilize the unimodal data better even when they are from different domains.

4.3 Unimodal Evaluation

We analyze how incorporating the multimodal correspondences into each unimodal encoder after alignment could benefit the downstream tasks. Figure 5 shows the accuracy for seismic and acoustic modalities before and after cross-modal alignment in the MOD application. With limited multimodal pairs, the pretrained unimodal encoders could gain the most significant performance improvements with InfoMAE. This emphasizes the InfoMAE's superior efficiency in enforcing cross-modal correspondence to each modality

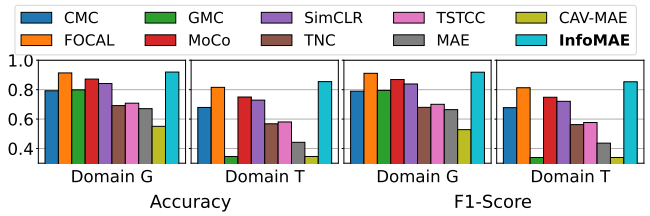

Figure 6: Performance of Joint Pretraining on MOD (seismic and acoustic) dataset and then finetuned on unseen domains.

to improve their downstream performance, with only a few multimodal pairs required. With InfoMAE, the aligned unimodal model can generate the most holistic representations through distributional alignment compared to geometric alignment (CMC, FOCAL).

4.4 Multimodal Pairing Efficiency

We also evaluate InfoMAE's alignment performance at varying amounts of multimodal data for MOD application in Table 2. We align both encoders pretrained from domain M ($M_{\text{sei}}||M_{\text{aco}}$) and compare them to standard joint pretraining with different ratios of multimodal data. Additionally, we provide supervised performance on the same amount of labeled data used for finetuning. InfoMAE consistently achieves superior multimodal data efficiency, with minimal degradation as we reduce the number of multimodal pairs. InfoMAE has an average of 3.42% gain over the highest-performing baselines and over 60% compared to joint model pretraining, which performs poorly in the absence of multimodal data. Joint pretraining even performs worse than the supervised approach with only 5% of multimodal data, indicating the standard self-supervised pretraining fails to learn effective representations with an insufficient amount of synchronized multimodal data. In contrast, the two-stage learning paradigm of InfoMAE leveraging widely available unsynchronized unimodal data could effectively mitigate this problem.

4.5 Standard Mutimodal Pretraining on Large-scale Synchronized Dataset

While InfoMAE excels as an efficient cross-modal alignment framework under limited pairs, it also demonstrates remarkable flexibility as a standard multimodal SSL framework. We evaluate InfoMAE against prior state-of-the-art works on Joint Multimodal Pretraining using abundant multimodal pairs, as shown in Figure 6. We use synchronized, unlabeled multimodal data from the MOD dataset to pretrain backbone encoders. Then we freeze the pretrained encoders and perform linear probing using labeled multimodal data

Table 4: Ablation accuracy of MOD cross-modal alignment.

Frameworks	$T_{\text{sei}}\|\|M_{\text{aco}}$	$G_{\text{sei}}\|\|T_{\text{aco}}$	$T_{\text{sei}}\|\|T_{\text{aco}}$	$G_{\text{sei}}\|\|M_{\text{aco}}$	$T_{\text{sei}}\|\|G_{\text{aco}}$
noTemp	0.6946	0.5881	0.5044	0.7435	0.6651
noShared	0.7683	0.6504	0.5298	0.8125	0.7395
noPrivate	0.5479	0.4180	0.2873	0.6259	0.5399
noAug	0.7863	0.6973	0.5881	0.8232	0.7924
InfoMAE	**0.7950**	**0.6986**	**0.5928**	**0.8326**	**0.8326**

from domains G and T, as described in Section 4.1. InfoMAE consistently outperforms the MAE-based framework and achieves better performance than other contrastive baselines. We leave more evaluation on Joint Multimodal Pretraining across four real-world datasets to Appendix E. Prior works, primarily designed for joint multimodal pretraining, often struggle with limited multimodal pairs and show significant performance degradation. In contrast, InfoMAE not only improves multimodal pairing efficiency but maintains high performance with minimal performance degradation.

4.6 Ablation Study

Finally, we study how each module of InfoMAE contributes to its performance through ablation studies. We evaluate four variants of InfoMAE by removing temporal, shared, private, and augmentation components in Table 4. The absence of either shared or private components leads to a significant degradation, implying the significance of factorized representation for cross-modal alignment. The drop in performance after removing temporal locality constraints also indicates the importance of learning temporal correspondence for time-series signals. Without temporal locality, the learned representations lose crucial temporal correspondence and can significantly compromise the ability to learn multimodal correspondences on top of the unimodal representations. Conversely, InfoMAE without augmentations does not significantly reduce the performance, demonstrating its robustness toward augmentation choices, in contrast to many contrastive learning frameworks that require careful selection of augmentations to avoid representational collapses.

5 Related Work

Self-Supervised Multimodal Learning. Self-supervised learning (SSL) techniques, such as Contrastive Learning (CL) and masked reconstructions, have achieved significant success in visual, textual, and time-series representation learning [5, 14, 15, 18, 55, 58, 63, 74, 76, 78]. Masked reconstruction learns informative representations by reconstructing masked inputs [4, 13, 23, 34, 73], with various masking strategies explored [2, 30, 77], and extended to time-frequency spectrograms [26, 29] and videos [19, 64]. Multimodal representation learning has become increasingly important with diverse applications [3, 38, 56, 57, 79]. Recent works leverage CL to learn correspondences between modalities [11, 51, 53, 54, 62, 66, 80], and others pretrain unified encoders for multimodal representations [25, 47]. Factorized Multimodal Learning [24, 28, 37, 40, 67] further decouples multimodal learning by acknowledging both modality-specific and modality-shared information. FOCAL [40] proposed contrastive learning objectives to learn shared and private representation in the orthogonal space. FactorizedCL [37] separates the shared and private space based on their relevance to the downstream tasks. Some works [17, 70] combine CL with MAE to capture cross-modal correspondence. Yet, these works minimize the geometric modality gap to learn cross-modal correspondences and rely on massive amounts of multimodal data for joint multimodal pretraining. In contrast, InfoMAE minimizes the information modality gap to further enhance the downstream performance. In reducing multimodal data pairs for training, many works [45, 65, 69] propose to impute missing modality pairs through feature generations. Wang *et al.* [71] proposes using CL to align multimodal encoders through an anchor modality yet still overlooking unimodal data. In contrast, InfoMAE minimizes the reliance on multimodal data by taking advantage of a large amount of unimodal data.

Multimodal Information Theory. There has been a long history of exploring common information between random variables in information theory [16, 72, 75], and it is still an active research field [20–22, 60]. However, it remains challenging to compute the common information in practical applications. Kleinman1 *et al.* [33] combines Variational Autoencoders with Gacs-Korner Common Information. Mai *et al.* [46] proposes to measure the information redundancy for multimodal data. However, they do not explicitly consider the unique information for factorization. InfoMAE adopts the informational factorization considering both private and shared information to construct a joint representation in a task-agnostic manner rather than extracting task-related information like [37].

6 Discussion & Conclusion

In this paper, we proposed InfoMAE, a pairing-efficient multi-stage SSL paradigm for multimodal IoT sensing. It first pretrains independent modality encoders on large-scale unimodal data sets. Then, it leverages a novel information theory-based optimization to achieve distributional cross-modal alignment with only limited multimodal pairs. Extensive evaluations compared to standard multimodal SSL frameworks demonstrated the superior efficiency and effectiveness of InfoMAE across multiple real-world IoT applications. We believe it opens new opportunities for developing more data-efficient and qualitative self-supervised multimodal models. In the Appendix, we provide additional evaluations and describe more details on the proof, datasets, implementation, and limitations.

Acknowledgments

Research reported in this paper was sponsored in part by the Army Research Laboratory under Cooperative Agreement W911NF-17-20196, NSF CNS 20-38817, and the Boeing Company. It was also supported in part by ACE, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the CCDC Army Research Laboratory, or the US government. The US government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

References

[1] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020.

[2] W. G. C. Bandara, N. Patel, A. Gholami, M. Nikkhah, M. Agrawal, and V. M. Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14507–14517, 2023.

[3] D. O. Bos et al. Eeg-based emotion recognition. *The Influence of Visual and Auditory Stimuli*, 56(3):1–17, 2006.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[5] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[8] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021.

[9] C.-Y. Chuang, R. D. Hjelm, X. Wang, V. Vineet, N. Joshi, A. Torralba, S. Jegelka, and Y. Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681, 2022.

[10] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[11] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. 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.

[12] J. DelPreto, C. Liu, Y. Luo, M. Foshey, Y. Li, A. Torralba, W. Matusik, and D. Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *Advances in Neural Information Processing Systems*, 35:13800–13813, 2022.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan. Time-series representation learning via temporal and contextual contrasting. In *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[15] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan. Self-supervised contrastive representation learning for semi-supervised time-series classification. *arXiv preprint arXiv:2208.06616*, 2022.

[16] P. Gacs and J. Körner. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2, 01 1973.

[17] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2022.

[18] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[19] A. Gupta, J. Wu, J. Deng, and F.-F. Li. Siamese masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] O. A. Hanna, X. Li, S. Diggavi, and C. Fragouli. Common information dimension. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 406–411. IEEE, 2023.

[21] O. A. Hanna, X. Li, S. Diggavi, and C. Fragouli. On the relation between the common information dimension and wyner common information. In *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024.

[22] O. A. Hanna, X. Li, C. Fragouli, and S. Diggavi. Can we break the dependency in distributed detection? In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2720–2725. IEEE, 2022.

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[24] W.-N. Hsu and J. Glass. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*, 2018.

[25] R. Hu and A. Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021.

[26] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

[27] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021.

[28] Q. Jiang, C. Chen, H. Zhao, L. Chen, Q. Ping, S. D. Tran, Y. Xu, B. Zeng, and T. Chilimbi. Understanding and constructing latent modality structures in multimodal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7661–7671, 2023.

[29] D. Kara, T. Kimura, Y. Chen, J. Li, R. Wang, Y. Chen, T. Wang, S. Liu, and T. Abdelzaher. Phymask: An adaptive masking paradigm for efficient self-supervised learning in iot. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, pages 97–111, 2024.

[30] D. Kara, T. Kimura, L. Shengzhong, L. Jinyang, L. Dongxin, W. Tianshi, W. Ruijie, C. Yizhuo, H. Yigong, and A. Tarek. Freqmae: Frequency-aware masked autoencoder for multi-modal iot sensing. In *The World Wide Web Conference*, 2024.

[31] H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, 2018.

[32] T. Kimura, J. Li, T. Wang, Y. Chen, R. Wang, D. Kara, M. Wigness, J. Bhattacharyya, M. Srivatsa, S. Liu, et al. Vibrofm: Towards micro foundation models for robust multimodal iot sensing. In *2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, pages 10–18. IEEE, 2024.

[33] M. Kleinman, A. Achille, S. Soatto, and J. Kao. Gacs-korner common information variational autoencoder. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] L. Kong, M. Q. Ma, G. Chen, E. P. Xing, Y. Chi, L.-P. Morency, and K. Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7918–7928, 2023.

[35] B. Korbar, D. Tran, and L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

[36] J. Li, Y. Chen, T. Kimura, T. Wang, R. Wang, D. Kara, Y. Hu, L. Wu, W. A. Hanafy, A. Souza, et al. Acies-os: A content-centric platform for edge ai twinning and orchestration. In *2024 33rd International Conference on Computer Communications and Networks (ICCCN)*, pages 1–9. IEEE, 2024.

[37] P. P. Liang, Z. Deng, M. Q. Ma, J. Y. Zou, L.-P. Morency, and R. Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36, 2023.

[38] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning.

[39] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[40] S. Liu, T. Kimura, D. Liu, R. Wang, J. Li, S. Diggavi, M. Srivastava, and T. Abdelzaher. Focal: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. *Advances in Neural Information Processing Systems*, 36, 2023.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021.

[42] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

[43] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[44] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

[45] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021.

[46] S. Mai, Y. Zeng, and H. Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 2022.

[47] D. Mizrahi, R. Bachmann, O. Kar, T. Yeo, M. Gao, A. Dehghan, and A. Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

[48] P. Morgado, Y. Li, and N. Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020.

[49] P. Morgado, I. Misra, and N. Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021.

[50] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[51] X. Ouyang, X. Shuai, J. Zhou, I. W. Shi, Z. Xie, G. Xing, and J. Huang. Cosmo: Contrastive fusion learning with small data for multimodal human activity recognition. In *International Conference on Mobile Computing And Networking (MobiCom)*, 2022.
[52] R. J. Piechocki, X. Wang, and M. J. Bocus. Multimodal sensor fusion in the latent representation space. *Scientific Reports*, 13(1):2005, 2023.
[53] N. Pielawski, E. Wetzer, J. Öfverstedt, J. Lu, C. Wählby, J. Lindblad, and N. Sladoje. Comir: Contrastive multimodal image representation for registration. *Advances in neural information processing systems*, 33:18433–18444, 2020.
[54] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning (ICML)*, 2022.
[55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
[56] A. Reiss and D. Stricker. Introducing a new benchmarked dataset for activity monitoring. In *International Symposium on Wearable Computers (ISWC)*, 2012.
[57] P. Schmidt, A. Reiss, R. Dürichen, C. Marberger, and K. V. Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *ICMI 2018*, pages 400–408. ACM, 2018.
[58] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer. How much can clip benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2021.
[59] M. Sugiyama, T. Suzuki, and T. Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012.
[60] E. Sula and M. Gastpar. The gray-wyner network and wyner's common information for gaussian sources. *IEEE Transactions on Information Theory*, 68(2):1369–1384, 2022.
[61] T. Sztyler and H. Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2016.
[62] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, 2020.
[63] S. Tonekaboni, D. Eytan, and A. Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations (ICLR)*, 2021.
[64] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
[65] L. Tran, X. Liu, J. Zhou, and R. Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414, 2017.
[66] D. J. Trosten, S. Lokse, R. Jenssen, and M. Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1255–1265, 2021.
[67] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2018.
[68] G. Verma, E. G. Dhekane, and T. Guha. Learning affective correspondence between music and image. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3975–3979. IEEE, 2019.
[69] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. 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*, pages 15878–15887, 2023.
[70] Y. Wang, J. Wang, B. Chen, Z. Zeng, and S.-T. Xia. Contrastive masked autoencoders for self-supervised video hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2733–2741, 2023.
[71] Z. Wang, Y. Zhao, H. Huang, J. Liu, A. Yin, L. Tang, L. Li, Y. Wang, Z. Zhang, and Z. Zhao. Connecting multi-modal contrastive representations. *Advances in Neural Information Processing Systems*, 36:22099–22114, 2023.
[72] A. Wyner. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, 21(2):163–179, 1975.
[73] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.
[74] Y. Yang, X. Liu, J. Wu, S. Borac, D. Katabi, M.-Z. Poh, and D. McDuff. Simper: Simple self-supervised learning of periodic targets. In *International Conference on Learning Representations*, 2023.
[75] L. Yu, V. Y. Tan, et al. Common information, noise stability, and their extensions. *Foundations and Trends® in Communications and Information Theory*, 19(2):107–389, 2022.
[76] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu. Ts2vec: Towards universal representation of time series. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.
[77] Q. Zhang, Y. Wang, and Y. Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022.
[78] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. In *Neural Information Processing Systems (NeurIPS)*, 2022.
[79] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong. Deep multisensor learning for missing-modality all-weather mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:254–264, 2021.
[80] M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459, 2021.

Appendix

A Information Formulation

A.1 Proof of the Equivalence between (1) and (2)

We first show the equivalence between the condition (1) and the constraints in (2) by proving the following proposition.

Proposition A.1. *For random variables X_1, X_2, if $U = s_1(X_1) = s_2(X_2)$, and there exists $W = g_1(X_1) = g_2(X_2)$ such that $X_1 \perp\!\!\!\perp X_2 \mid W$, then the following two statements are equivalent.*

(a) $(\forall f_1, f_2)\Big([f_1(X_1) = f_2(X_2)] \implies [(\exists f) f(U) = f_1(X_1) = f_2(X_2)]\Big)$.

(b) There is a one-to-one mapping between W and U (i.e., $X_1 \perp\!\!\!\perp X_2 \mid U$).

Proof. We first prove the direction (b) $\implies$ (a) using properties of basic information-theory measures (Chapter 2 in [10]). For any f_1, f_2 such that $f_1(X_1) = f_2(X_2)$, we have

$$0 \overset{(i)}{=} I(X_1; X_2|U) \overset{(ii)}{\geq} I(f_1(X_1); f_2(X_2)|U) \overset{(iii)}{\geq} 0, \tag{12}$$

where (i) follows that X_1 and X_2 are independent conditioned on U; (ii) is due to the data processing inequality of mutual information; and (iii) is because the mutual information is always non-negative. (12) implies that $I(f_1(X_1); f_2(X_2)|U) = 0$. In addition, since $I(f_1(X_1); f_2(X_2)|U) = H(f_1(X_1)|U) - H(f_1(X_1)|f_2(X_2), U)$ and $H(f_1(X_1)|f_2(X_2), U) = 0$, we have $H(f_1(X_1)|U) = 0$. This concludes that there exist a deterministic function f such that $f(U) = f_1(X_1) = f_2(X_2)$.

Next, we prove the other direction (a) $\implies$ (b). Note that W given in the proposition statement satisfies $W = g_1(X_1) = g_2(X_2)$ and therefore, from (a), we know that there exist a function h_1 such that $W = h_1(U)$. Since W also satisfies that $X_1 \perp\!\!\!\perp X_2 \mid W$ and $U = s_1(X_1) = s_2(X_2)$, then applying the direction (b) $\implies$ (a), we have that $U = h_2(W)$ for some function h_2. Therefore, there is a one-to-one mapping between W and U. □

Note that it is difficult to obtain a random variable U that satisfies (a) (i.e. the sufficient common variable in Defined 2.2). The Proposition A.1 allows us to find a random variable W (if it exists) instead. And the one with minimum entropy can be obtained by solving the optimization problem (2).

A.2 Derivation of the Shared Loss (4)

We first group the terms that only depend on U_1 or U_2 as follows.

$$\begin{aligned}\mathcal{L}_{\text{info}}^{\text{shared}} &= \alpha d(U_1, U_2) + \beta(H(U_1) + H(U_2)) + I(X_1; X_2 \mid U_1) \qquad (13)\\ &\quad + I(X_1; X_2 \mid U_2)\\ &= \alpha d(U_1, U_2) + \mathcal{L}(U_1) + \mathcal{L}(U_2), \qquad (14)\end{aligned}$$

Table 5: Statistical summaries of domains and datasets

Dataset	Modalities (Freq)	Sample Length	Overlap	Classes	#Pretrain Samples	Used for Alignment	#Alignment Samples	# Finetune Samples
Domain M	acoustic (8kHz) seismic (100Hz)	2 sec	0%	5 sec	39,609	✓	1981	734
Domain G	acoustic (8kHz) seismic (100Hz)	2 sec	0%	2 sec	35,168	✗	-	3136 (joint)
Domain T	acoustic (8kHz) seismic (100Hz)	2 sec	0%	4	43,819	✗	-	4205 (joint)
PAMAP2	acc, gyro, mag, lig (all 50Hz)	5 sec	50%	18	9,611	✓	4805	961
RW-HAR	acc, gyr, mag (all 100Hz)	2 sec	50%	8	12,887	✗	-	-

where $d(U_1, U_2)$ can be measured using the Euclidean distance or other distance measures. And

$$\begin{aligned}
\mathcal{L}(U_1) &= I(X_1; X_2|U_1) + \beta H(U_1) \\
&\overset{(i)}{=} I(X_1; X_2|U_1) + \beta I(X_1; U_1) \\
&\overset{(ii)}{=} \mathbb{E}_{U_1}\left[D_{KL}(p_{X_1,X_2|U_1}||p_{X_1|U_1}p_{X_2|U_1})\right] \\
&\quad + \beta D_{KL}(p_{X_1,U_1}||p_{X_1}p_{U_1}) \qquad (15)\\
&= \mathbb{E}_{X_1,X_2,U_1}\left[\log\frac{p_{X_1,X_2|U_1}}{p_{X_1|U_1}p_{X_2|U_1}}\right] + \beta\mathbb{E}_{X_1,U_1}\left[\log\frac{p_{X_1,U_1}}{p_{X_1}p_{U_1}}\right] \\
&= \mathbb{E}_{X_1,X_2,U_1}\left[\log\frac{p_{X_1,X_2,U_1}p_{U_1}}{p_{X_1,U_1}p_{X_2,U_1}}\right] + \beta\mathbb{E}_{X_1,U_1}\left[\log\frac{p_{X_1,U_1}}{p_{X_1}p_{U_1}}\right] \\
&= \mathbb{E}_{X_1,X_2,U_1}\left[\log\frac{p_{X_1,X_2,U_1}}{p_{X_1}p_{X_2}p_{U_1}} + \log\frac{p_{X_1}p_{U_1}}{p_{X_1,U_1}} + \log\frac{p_{X_2}p_{U_1}}{p_{X_2,U_1}}\right] \qquad (16)\\
&\quad + \beta\mathbb{E}_{X_1,U_1}\left[\log\frac{p_{X_1,U_1}}{p_{X_1}p_{U_1}}\right] \\
&= \mathbb{E}_{X_1,X_2,U_1}\left[\log\frac{p_{X_1,X_2,U_1}}{p_{X_1}p_{X_2}p_{U_1}}\right. \qquad (17)\\
&\quad \left. +(1-\beta)\log\frac{p_{X_1,U_1}}{p_{X_1}p_{U_1}} + \log\frac{p_{X_2,U_1}}{p_{X_2}p_{U_1}}\right], \qquad (18)
\end{aligned}$$

where (i) follows the relation between mutual information an entropy that $I(X_1; U_1) = H(U_1) - H(U_1|X_1)$ and $H(U_1|X_1) = 0$ because U_1 is a deterministic function of X_1; (ii) is by definition of the conditional mutual information; and the remaining equalities use the Bayes' rule. Similarly, we have

$$\begin{aligned}
&\mathcal{L}(U_2) = I(X_1; X_2|U_2) + (1-\beta)H(U_2) \\
&= \mathbb{E}_{X_1,X_2,U_2}\left[\log\frac{p_{X_1,X_2,U_2}}{p_{X_1}p_{X_2}p_{U_2}} + \log\frac{p_{X_1,U_2}}{p_{X_1}p_{U_2}} + \beta\log\frac{p_{X_2,U_2}}{p_{X_2}p_{U_2}}\right] \qquad (19)
\end{aligned}$$

Combining (14), (18) and (19), we can obtain

$$\begin{aligned}
\mathcal{L}_{\text{info}}^{\text{shared}} &= \alpha d(U_1, U_2) + \sum_{i=1}^{2}\mathbb{E}_{X_1,X_2,U_i}\left[\log\frac{p_{X_1,X_2,U_i}}{p_{X_1}p_{X_2}p_{U_i}}\right. \qquad (20)\\
&\quad \left. +(1-\beta)\log\frac{p_{X_i,U_i}}{p_{X_i}p_{U_i}} + \log\frac{p_{X_{3-i},U_i}}{p_{X_{3-i}}p_{U_i}}\right]. \qquad (21)
\end{aligned}$$

A.3 Derivation of the Private Loss (7)

Similar to (18), since $H(V_1|X_1) = H(V_2|X_2) = 0$, we have that

$$\begin{aligned}
\mathcal{L}_{\text{info}}^{\text{private}} &= \gamma H(V_1) + \gamma H(V_2) + \epsilon I(V_1; U_1) + \epsilon I(V_2; U_2), \\
&= \gamma I(X_1; V_1) + \gamma I(X_2; V_2) + \epsilon I(V_1; U_1) + \epsilon I(V_2; U_2), \\
&= \sum_i \mathbb{E}_{X_i,V_i,U_i}\left[\gamma\log\frac{p_{X_i,V_i}}{p_{X_i}p_{V_i}} + \epsilon\log\frac{p_{V_i,U_i}}{p_{V_i}p_{U_i}}\right].
\end{aligned} \qquad (22)$$

B Datasets

This section describes the cross-modal alignment and joint multimodal pretraining datasets from two applications: Moving Object Detection (MOD) and Human Activity Recognition (HAR). Table 5 provides the statistical values of each domain.

B.1 Cross-modal Alignment Datasets

B.1.1 Moving Object Detection. We have seismic and acoustic signals describing different vehicles on three different domains. For simplicity, we use one letter to represent each domain.

Domain M is a publicly released [40] moving object detection dataset consisting of signals from 7 different moving vehicles, recorded at three different distances and four different speeds.

Domain G contains a self-collected dataset on state park grounds near an outdoor research facility with four sensor nodes deployed. The dataset contains four distinct targets navigating the neighborhood near the sensors in some arbitrary order.

Domain T has a similar setup as MOD but involves different targets and scenes. This set contains data collected from a paved parking lot, unpaved trails, and gravel roads within a park. Vibration signals of 2 standard-size SUVs from different manufacturers, one lightweight sports car, and one muscle car were recorded. One hour of data for each vehicle was collected at each scene. use the first 50 minutes for training and the last 10 minutes for validation and testing.

B.1.2 Human Activity Recognition. Unlike the MOD application, where we used data from different domains for unimodal pretraining, we leveraged two different HAR datasets for unimodal pretraining and cross-modal alignment to evaluate the scenario in which IMU data has high degrees of heterogeneity.

RW-HAR [61] is a public dataset with accelerometer, gyroscope, magnetometer, and light signals sampled at 50Hz. It includes data from 15 subjects performing 8 human activities. We use the data collected from the subjects' waist and randomly select ten subjects for training, 2 for validation, and 3 for testing.

PAMAP2 [56] contains inertial data from 18 human daily activities performed by 9 subjects. PAMAP2 includes 9,611 instances, with data captured using inertial measurement units (IMUs) placed on the chest, the wrist of the dominant arm, and the dominant side's ankle. We use the data collected from the wrist. The signal is collected at a sampling rate of 100Hz. 7 random subjects are used for training, and 2 subjects for testing.

Combined is a concatenated dataset of RealWord-HAR and PAMAP2. Since PAMAP2 does not contain any light signals, we drop the light modality and only use the three IMU modalities for evaluation.

Table 6: Inference profiling on Raspberry Pi 4 device.

App.	P99 (s)	Average (s)	Model Size (MB)	# Parameters (M)
MOD	0.5803	0.2259	47.9820	12.565831
HAR	0.1728	0.1690	17.8810	4.669818

Table 7: Cross-modal alignment with sparse pairs.

Framework	GMC		FOCAL		InfoMAE	
Ratio	Acc	F1	Acc	F1	Acc	F1
0.01	0.8252	0.8247	0.8573	0.8556	**0.8794**	**0.8786**
0.02	0.8305	0.8272	0.8580	0.8573	**0.8821**	**0.8811**
0.03	0.8667	0.865	0.8560	0.8529	**0.8875**	**0.8841**

C Data Preprocessing

We partition the time-series data into segments of uniform length. Each segment is subdivided into intervals with overlaps. We apply the Fourier transform to the signal in each interval to derive its spectral content, thereby retaining both temporal and spectral characteristics. During training, we adopt the same augmentations as FOCAL [40] to the input before and after the fourier-transform.

D Experiment and Implementation Details

During pretraining, we randomly sample a batch of sequences of L consecutive samples. We jointly optimize the backbone encoders and decoders with AdamW [43] optimizer and Cosine scheduler [42]. We also train discriminators for density-ratio estimations [31, 59]. We apply convolution blocks to map the time-frequency sample into a one-dimensional embedding to match the input dimension X_1 with their shared and private representations V_1, U_1, followed by 5-layer MLP to their density ratio.

While InfoMAE's training requires additional computation due to the discriminators and the MAE architecture, we would like to note that InfoMAE incurs no extra inference overhead. We evaluated InfoMAE's inference performance on a Raspberry Pi 4 device and present the computational overhead in Table 6. The result demonstrates that InfoMAE achieves real-time inference in less than 1 second, making it suitable for real-time deployment in WoT/IoT applications where only low-end devices are available.

Computation. We conducted our experiments on NVIDIA RTX 4090 GPUs (24GB). The training time varies from minutes for finetuning to 2 days for pretraining. The training time for cross-modal alignment is faster with fewer multimodal pairs.

E Additional Evaluation

E.1 Evaluation: Sparse cross-modal alignment

We conduct additional experiments to evaluate InfoMAE under extremely sparse conditions, reducing the availability of multimodal pairs to as low as 1%, 2%, and 3%. These results, presented in Table 7, highlight that InfoMAE continues to outperform top-performing baselines across these extremely constrained scenarios. The findings illustrate InfoMAE's robustness in aligning representations under sparse multimodal pairing conditions.

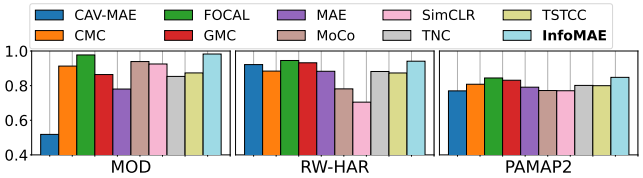

Figure 7: Joint Multimodal Pretraining compared with previous joint pretraining SSL frameworks on four datasets.

Table 8: Ablation accuracy with Joint Pretraining.

Frameworks	MOD	RW-HAR	PAMAP2
woTemp	0.8734	0.8442	0.6948
woShared	0.9531	0.8771	0.8095
woPrivate	0.9082	0.9100	0.8080
woAugmentation	0.9538	0.9106	0.8163
InfoMAE	**0.9826**	**0.9411**	**0.8478**

E.2 Joint Multimodal Pretraining

Although InfoMAE is primarily designed for learning settings where the multimodal pairs are scarce, InfoMAE demonstrates strong flexibility and generalization as a standard multimodal SSL framework when abundant multimodal pairs are available. Figure 7 presents additional finetuning performance on joint multimodal pretraining. InfoMAE significantly exceeds the MAE-based SSL framework and achieves comparable or superior performance to the SOTA baselines. It is noteworthy that these baselines are mainly designed for joint multimodal pretraining. InfoMAE is a universal framework for cross-modal alignment that achieves comparable performance as multimodal SSL with few sacrifices.

E.3 Additional Ablation Studies

In Table 8, we present additional ablation accuracy on joint multimodal pretraining, evaluating variants InfoMAE when abundant multimodal data is available. We find the results consistent with the performance presented in Section 4.6

F Limitations and Future Work

Pretraining Overhead and Efficiency. Compared to contrastive SSL (*e.g.*, FOCAL, CMC, etc.), InfoMAE incurs additional computational overhead due to its autoencoder architecture and density ratio estimation. While this enhances multimodal alignment, it increases training complexity. Future work could explore concurrent unimodal pretraining, optimized attention mechanisms like FlashAttention, and alternative density ratio estimation techniques without training discriminators to improve efficiency.

Potential Bias and Robustness Under Sparse Sampling. InfoMAE demonstrates resilience under sparse multimodal settings (Appendix E.1). However, we would like to note that distribution-based alignment cannot completely eliminate sampling biases, which can affect learned representations. Further research is required to develop more robust alignment methods that mitigate sampling errors and improve generalization under extreme data sparsity.