# CORRCLR: CORRELATION-BASED CONTRASTIVE LEARNING OF REPRESENTATIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Contrastive learning (CL) is a central self-supervised paradigm that has driven major advances in multimodal representation learning. However, its performance is highly sensitive to the temperature hyperparameter in loss functions such as InfoNCE, and the literature lacks a systematic approach that generalizes across diverse settings. In this work, we establish a fundamental relationship between the temperature hyperparameter and the correlation coefficient of multimodal representations. Leveraging this insight, we reformulate contrastive loss functions and introduce their correlation-based counterparts. While temperature hyperparameters are difficult to tune, correlation coefficients provide an interpretable and measurable alternative. Consequently, we propose CorrCLR, a novel algorithm that learns multimodal representations by dynamically updating correlation coefficients as the representations evolve. Experiments across diverse multimodal settings, including 3D point-image-text, video-audio-text, and mmWave-WiFi-RFID, demonstrate that CorrCLR delivers robust performance improvements over state-of-the-art baselines across various downstream tasks, such as zero-shot classification and retrieval.

## 1 INTRODUCTION

Real-world environments are inherently multimodal, and humans naturally integrate diverse sensory inputs, such as visual, tactile, and auditory signals, to form a holistic understanding of their surroundings (Xiao et al., 2024; Dufumier et al., 2025). This capability makes multimodal representation learning a crucial paradigm for developing intelligent systems that can align information across modalities and enhance generalization (Xia et al., 2023; Tjandrasuwita et al., 2025; Yin et al., 2026). Contrastive learning (CL), a prominent self-supervised technique, has enabled significant advances in this area by aligning representations from different modalities within a shared latent space (Radford et al., 2021; Li et al., 2021; Chen et al., 2023b; Xue et al., 2024). While InfoNCE (Oord et al., 2019) is a widely used loss function for dual-modality alignment, scaling to more than two modalities often relies on pairwise alignment (Girdhar et al., 2023; Liu et al., 2023; Zhu et al., 2024). However, such pairwise approaches fail to effectively capture higher-order dependencies among multimodal representations (Koromilas et al., 2025). Several recent works have proposed various contrastive loss functions for aligning more than two modalities (Cicchetti et al., 2025b; Saporta et al., 2024; Cicchetti et al., 2025a). For example, Muscle loss was recently proposed by Setayesh et al. (2026) as a principled extension of InfoNCE that enables the alignment of multimodal representations while more effectively capturing their underlying dependencies.

Despite these advances, contrastive loss functions such as InfoNCE and Muscle critically depend on temperature hyperparameters that control the sharpness of the similarity distributions (Wang & Liu, 2021; Huang et al., 2023; Kim & Kim, 2025). In multimodal settings, this introduces an expanding set of parameters whose manual tuning becomes increasingly intractable as the number of modalities grows. Existing literature offers limited guidance: while some methods treat temperature as a learnable parameter (Radford et al., 2021; Qiu et al., 2023), others advocate for fixed values (Chen et al., 2020a; Girdhar et al., 2023), often leading to conflicting empirical findings (Zhu et al., 2024). Furthermore, current adaptive temperature strategies are primarily designed either for SimCLR-style multiview settings (Chen et al., 2020a), as in MACL (Huang et al., 2023) and DySTreSS (Manna et al., 2025), or for dual-modality settings, as in iSogCLR (Qiu et al., 2023). Thus, they do not natu-



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.

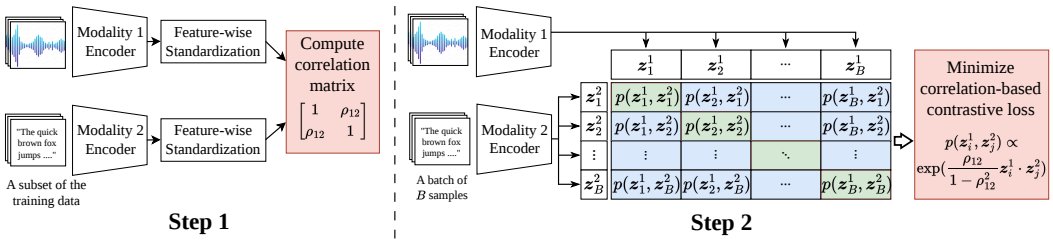



Figure 1: Overview of the CorrCLR framework. Step 1: Compute representations using modality encoders and update the correlation matrix from standardized representations. Step 2: Update modality encoders by minimizing the correlation-based contrastive loss given the updated correlation matrix.

rally generalize to scenarios with more than two modalities. Consequently, a unified and systematic approach to temperature control in multimodal CL remains absent.

Our central idea is to replace heuristic temperature parameters with interpretable and measurable alternatives. Recently, Betser et al. (2026) theoretically showed that InfoNCE induces a Gaussian structure in representations. This observation is not coincidental. In particular, both the InfoNCE and Muscle loss functions are derived from density ratios over representations (Oord et al., 2019; Setayesh et al., 2026). That is, the general formulation of InfoNCE can be viewed as maximizing the joint probability of positive sample representations while minimizing that of negative samples (see Step 2 in Fig. 1). InfoNCE assumes that the density ratio over representations is proportional to the exponential of their cosine similarity divided by a temperature parameter (Oord et al., 2019). By modeling the joint probability distribution of representations as a multivariate Gaussian, we provide a mathematical basis for this assumption. Consequently, we establish a fundamental relationship between temperature parameters and the correlation coefficients of multimodal representations.

In particular, we demonstrate that, under the multivariate Gaussian assumption, the temperature parameter $\tau$ in InfoNCE is related to the correlation coefficient $\rho$ between the representations of two modalities via the relationship $\tau = \frac{1-\rho^2}{\rho}$. This implies that a higher $\rho$ corresponds to a lower $\tau$, and vice versa. We further show that the temperature parameters in Muscle loss are related to the elements of the inverse correlation matrix. We reformulate the InfoNCE and Muscle loss functions by replacing temperature parameters with correlation coefficients, thereby introducing $\rho$-InfoNCE and $\rho$-Muscle. Building on this, we propose CorrCLR (**Corr**elation-based **C**ontrastive **L**earning of **R**epresentations), which learns both representations and correlation coefficients in an iterative manner. As shown in Fig. 1, CorrCLR consists of two primary steps: in Step 1, we update the correlation matrix using representations obtained from the modality encoders; in Step 2, we update the encoders by minimizing the correlation-based contrastive loss given the updated correlation matrix.

In summary: (1) we establish a principled mathematical connection between the temperature parameters in the InfoNCE and Muscle loss functions and the correlation coefficients of multimodal representations; (2) we introduce $\rho$-InfoNCE and $\rho$-Muscle, which replace heuristic temperature parameters with interpretable and measurable correlation coefficients; (3) we propose CorrCLR, a novel algorithm that learns multimodal representations while adaptively updating their correlation coefficients; and (4) we empirically validate the effectiveness of CorrCLR across diverse settings, including 3D point-image-text, video-audio-text, and mmWave-WiFi-RFID, demonstrating consistent improvements over state-of-the-art baselines across various downstream tasks.

## 2 Related Work

**Multimodal contrastive learning** Multimodal learning aims to derive robust representations by leveraging complementary information across diverse modalities to enhance performance on downstream tasks (Ngiam et al., 2011; Huang et al., 2021). One possible approach is to employ CL to align multimodal representations within a shared latent space (Jia et al., 2021; Zhou et al., 2026). It can serve as a pretraining phase prior to fusion (Yang et al., 2023; Shi et al., 2025) and has also demonstrated significant success in enabling zero-shot capabilities (Chen et al., 2020a; Radford et al., 2021). InfoNCE (Oord et al., 2019) is the most widely adopted contrastive loss and was famously utilized in CLIP (Radford et al., 2021) to align image and text modalities. For settings involving more than two modalities, existing works often utilize pairwise alignment by applying InfoNCE to each pair. Notable examples include CLIP4VLA (Ruan et al., 2023), ULIP (Xue et al.,

2023), ImageBind (Girdhar et al., 2023), and LanguageBind (Zhu et al., 2024). However, pairwise alignment has been shown to be insufficient for fully capturing the dependencies among multimodal representations (Koromilas et al., 2025). To address this, various contrastive loss functions, such as Symile (Saporta et al., 2024), GRAM (Cicchetti et al., 2025b), and TRIANGLE (Cicchetti et al., 2025a), have been proposed to align representations across more than two modalities. Additionally, Muscle loss was recently proposed by Setayesh et al. (2026) as a principled extension of InfoNCE to more than two modalities, emphasizing the role of temperature parameters for each pair of modalities. Both InfoNCE and Muscle loss rely heavily on temperature hyperparameters. Despite the fact that model performance is highly sensitive to these parameters, the literature lacks a unified or principled guideline for their selection that generalizes across diverse settings.

**Temperature Tuning** Several studies utilizing InfoNCE or its variants rely on heuristic searches to determine an optimal fixed temperature. SimCLR (Chen et al., 2020a), MoCo (He et al., 2020), and MSC (Reiss & Hoshen, 2023) adopt temperature values of 0.1, 0.07, and 0.25, respectively. Alternatively, some approaches follow CLIP (Radford et al., 2021) by treating temperature as a learnable parameter (Liu et al., 2023; Xue et al., 2024; Zhu et al., 2024). However, ImageBind (Girdhar et al., 2023) empirically shows that a learnable temperature does not consistently outperform a carefully tuned fixed value. Additionally, some adaptive temperature strategies, such as MACL (Huang et al., 2023), TS (Kukleva et al., 2023), and DySTreSS (Manna et al., 2025), rely on the gradient structure of the InfoNCE loss within a multiview learning framework, while others, such as iSogCLR (Qiu et al., 2023) and TempNet (Qiu et al., 2024), are optimization-based temperature tuning algorithms designed for dual-modality settings. Thus, they may not be effectively generalized to multimodal settings. In this work, we establish a fundamental connection between the temperature hyperparameters in the InfoNCE and Muscle loss functions and the correlation coefficients of multimodal representations. While the importance of cross-modal correlation has recently been studied in multimodal sentiment analysis to improve fusion (Li et al., 2025; Xu et al., 2025) or address missing-modality issues (Zhuang et al., 2025), our work is the first to link temperature in contrastive loss functions to correlation coefficients and propose a practical algorithm for their estimation. We achieve this by modeling the joint probability distribution of multimodal representations as a multivariate Gaussian.

**InfoNCE and Gaussian Distribution** The inherent connection between the InfoNCE loss and the Gaussian distribution has been the focus of several recent studies. For instance, Lu et al. (2023) notes that among popular similarity functions for InfoNCE, the Gaussian kernel tends to yield superior empirical results. Furthermore, Bansal et al. (2025) theoretically demonstrates that when data samples are drawn from Gaussian mixture models, InfoNCE can identify the optimal lower-dimensional subspace and match the performance of fully supervised algorithms. Most recently, Betser et al. (2026) showed that high-dimensional representations learned via InfoNCE asymptotically approach a multivariate Gaussian distribution. By explicitly modeling the joint probability distribution of representations as a multivariate Gaussian, we provide a mathematical basis for the InfoNCE structure, in which the density ratio over representations is proportional to the exponential of their cosine similarity divided by a temperature parameter.

## 3 Correlation Coefficient-based Contrastive Loss Formulation

We first reformulate the InfoNCE loss function (Oord et al., 2019) and introduce $\rho$-InfoNCE. We show how the temperature parameter in InfoNCE is related to the correlation coefficient $\rho$ between the representations of two modalities. Then, we introduce $\rho$-Muscle, which is a reformulation of the Muscle loss function (Setayesh et al., 2026). In particular, we show how the correlation coefficient matrix among the representations of $M$ modalities replaces the temperature parameters used in the Muscle loss function.

### 3.1 Aligning Two Modalities

InfoNCE is a popular contrastive loss function that has been widely used to align the representations of two modalities (Radford et al., 2021; Girdhar et al., 2023). Let $\{\boldsymbol{z}_i^n\}_{i=1}^B$ and $\{\boldsymbol{z}_i^m\}_{i=1}^B$ denote the representations for a batch of $B$ samples from modalities $n$ and $m$, respectively. Taking $\boldsymbol{z}_i^n$ as the anchor, the InfoNCE loss is defined as follows (Oord et al., 2019):

$$\mathcal{L}_{\text{InfoNCE}}^{n,m}(\boldsymbol{z}_i^n) = -\log \frac{f\left(\boldsymbol{z}_i^n,\, \boldsymbol{z}_i^m\right)}{\sum_{j\in[B]} f\left(\boldsymbol{z}_i^n,\, \boldsymbol{z}_j^m\right)}, \tag{1}$$

where $f$ is a function that assigns high values to positive pairs and low values to negative pairs. A common choice for this function is $\exp(\boldsymbol{z}_i^n \cdot \boldsymbol{z}_j^m/\tau_{nm})$, where $\tau_{nm}$ is a temperature parameter that

moderates the similarity. Thus, the InfoNCE loss is formulated as follows (Oord et al., 2019):

$$\mathcal{L}^{n,m}_{\text{InfoNCE}}(\boldsymbol{z}^n_i) = -\log \frac{\exp\left(\boldsymbol{z}^n_i \cdot \boldsymbol{z}^m_i / \tau_{nm}\right)}{\sum_{j\in[B]} \exp\left(\boldsymbol{z}^n_i \cdot \boldsymbol{z}^m_j / \tau_{nm}\right)}. \tag{2}$$

In existing works, $\tau_{nm}$ in (2) is treated as a hyperparameter whose value varies across studies and significantly affects CL performance (Kim & Kim, 2025). However, we show that there is a relationship between $\tau_{nm}$ and the correlation coefficient between the representations of modalities $n$ and $m$. We reformulate the InfoNCE loss based on this new perspective and introduce $\rho$-InfoNCE.

As shown by Oord et al. (2019), the optimal value of $f(\boldsymbol{z}^n_i, \boldsymbol{z}^m_j)$ in (1) is proportional to the density ratio $p(\boldsymbol{z}^n_i, \boldsymbol{z}^m_j)/p(\boldsymbol{z}^n_i)p(\boldsymbol{z}^m_j)$. Thus, we have $p(\boldsymbol{z}^n_i, \boldsymbol{z}^m_j)/p(\boldsymbol{z}^n_i)p(\boldsymbol{z}^m_j) \propto \exp(\boldsymbol{z}^n_i \cdot \boldsymbol{z}^m_j/\tau_{nm})$. This proportionality implies that the density ratio of the representations is modeled as an exponential function of their similarity. Following this observation, let us assume that the joint probability density of the concatenated representations of modalities $n$ and $m$ is a multivariate Gaussian distribution with mean $\mathbf{0} \in \mathbb{R}^{2d}$ and covariance matrix $\boldsymbol{\Sigma} = \left[\begin{smallmatrix} 1 & \rho_{nm} \\ \rho_{nm} & 1 \end{smallmatrix}\right] \otimes \boldsymbol{I}_d \in \mathbb{R}^{2d\times 2d}$, where $d$ is the dimension of the representation vectors and $\rho_{nm} \in [-1, 1]$ is the correlation coefficient between the representations of modalities $n$ and $m$. The Gaussian assumption is consistent with prior studies showing that InfoNCE induces a Gaussian distribution (Betser et al., 2026; Klindt et al., 2026), while the considered covariance matrix structure is consistent with the InfoNCE formulation in (2), which considers only a single temperature parameter. Furthermore, $\rho_{nm}$ can be negative when the representations of two modalities are negatively correlated. However, in dual-modality settings, the sign of $\rho_{nm}$ does not affect the overall performance, as the modality encoders (i.e., neural networks) can adjust the signs of the representations accordingly. In Appendix A, we show that under the multivariate Gaussian distribution assumption, the proportionality $p(\boldsymbol{z}^n_i, \boldsymbol{z}^m_j)/p(\boldsymbol{z}^n_i)p(\boldsymbol{z}^m_j) \propto \exp(\frac{\rho_{nm}}{1-\rho^2_{nm}} \boldsymbol{z}^n_i \cdot \boldsymbol{z}^m_j)$ holds. This indicates that, consistent with the derivation of the InfoNCE loss function, we have modeled the density ratio of the representations as an exponential function of their similarity. However, because our formulation is rooted in an actual probability distribution, it provides deeper insights into the dynamics of multimodal CL. In particular, based on this formulation, there is a relationship between the temperature parameter $\tau_{nm}$ in the InfoNCE loss and the correlation coefficient between the representations of modalities $n$ and $m$ as follows: $\tau_{nm} = \frac{1-\rho^2_{nm}}{\rho_{nm}}$. Thus, a higher $\rho_{nm}$ between the representations of the modalities is equivalent to a lower $\tau_{nm}$, and vice versa. In Section 4, we show how this connection to the correlation coefficient can facilitate the CL of representations by estimating the correlation value instead of heuristically tuning the temperature parameter.

By setting $f(\boldsymbol{z}^n_i, \boldsymbol{z}^m_j)$ in (1) to $\exp(\frac{\rho_{nm}}{1-\rho^2_{nm}} \boldsymbol{z}^n_i \cdot \boldsymbol{z}^m_j)$, we define $\rho$-InfoNCE as follows:

$$\mathcal{L}^{n,m}_{\rho\text{-InfoNCE}}(\boldsymbol{z}^n_i) = -\log \frac{\exp\left(\frac{\rho_{nm}}{1-\rho^2_{nm}} \boldsymbol{z}^n_i \cdot \boldsymbol{z}^m_i\right)}{\sum_{j\in[B]} \exp\left(\frac{\rho_{nm}}{1-\rho^2_{nm}} \boldsymbol{z}^n_i \cdot \boldsymbol{z}^m_j\right)}. \tag{3}$$

Next, we extend our correlation-based contrastive loss formulation to more than two modalities.

### 3.2 Aligning Multiple Modalities

Recently, the Muscle loss has been proposed by Setayesh et al. (2026) to jointly align representations from multiple modalities. The Muscle loss can be considered a principled extension of InfoNCE to more than two modalities and, compared to pairwise alignment approaches, can effectively capture dependencies among all multimodal representations. Let $M$ denote the number of modalities. Taking $\boldsymbol{z}^n_i$, $n \in [M]$ as the anchor, the Muscle loss is defined as follows (Setayesh et al., 2026):

$$\mathcal{L}^n_{\text{Muscle}}(\boldsymbol{z}^n_i) = -\log \frac{f(\boldsymbol{z}^n_i, \{\boldsymbol{z}^m_i\}^M_{m=1,\, m\neq n})}{\sum_{\boldsymbol{j}\in\mathcal{J}^n} f(\boldsymbol{z}^n_i, \{\boldsymbol{z}^m_{j_m}\}^M_{m=1,\, m\neq n})}, \tag{4}$$

where $f$ is a function that assigns high values to positive $M$-tuples and low values to negative $M$-tuples. Specifically, for a given anchor $\boldsymbol{z}^n_i$, the set of representation vectors $\{\boldsymbol{z}^m_i\}^M_{m=1,\, m\neq n}$ constitutes a positive sample. Conversely, any combination of representation vectors, one from each modality in $[M] \setminus \{n\}$, is considered a negative sample if at least one vector in the combination corresponds to an index $j \neq i$. Additionally, we have $\mathcal{J}^n = \{\boldsymbol{j} = (j_1, \ldots, j_M) \mid j_m \in [B],\, m \in [M] \setminus \{n\}\}$. To remain consistent with

the InfoNCE loss for the two-modality case, it was shown by Setayesh et al. (2026) that $f(\boldsymbol{z}_i^n, \{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M)$ in (4) can be defined as $\alpha_{\boldsymbol{j}} \exp(\boldsymbol{z}_i^n \cdot \sum_{m\in[M]\setminus\{n\}} \boldsymbol{z}_{j_m}^m / \tau_{nm}^{(M)})$, where $\alpha_{\boldsymbol{j}} = \exp(-\frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\sum_{m'\in[M]\setminus\{n,m\}} \gamma_{mm'}^{(M)} \boldsymbol{z}_{j_m}^m \cdot \boldsymbol{z}_{j_{m'}}^{m'})$ and $\gamma_{mm'}^{(M)} = 1/\tau_{mm'}^{(M-1)} - 1/\tau_{mm'}^{(M)}$. Here, $\tau_{mm'}^{(M)}$ and $\tau_{mm'}^{(M-1)}$ are temperature parameters that moderate the similarity between the representations of modalities $m$ and $m'$ when $M$ and $M-1$ modalities are available, respectively. Thus, the Muscle loss is formulated as follows (Setayesh et al., 2026):

$$\mathcal{L}_{\text{Muscle}}^n(\boldsymbol{z}_i^n) = -\log \frac{\alpha_{(i,\ldots,i)} \exp(\boldsymbol{z}_i^n \cdot \sum_{m\in[M]\setminus\{n\}} \boldsymbol{z}_i^m / \tau_{nm}^{(M)})}{\sum_{\boldsymbol{j}\in\mathcal{J}^n} \alpha_{\boldsymbol{j}} \exp(\boldsymbol{z}_i^n \cdot \sum_{m\in[M]\setminus\{n\}} \boldsymbol{z}_{j_m}^m / \tau_{nm}^{(M)})}. \tag{5}$$

$\mathcal{L}_{\text{Muscle}}^n(\boldsymbol{z}_i^n)$ involves $(M-1)^2$ temperature parameters. Tuning this many hyperparameters poses a significant challenge. To address this, we propose a correlation-based formulation of the Muscle loss, which replaces these temperature parameters with the correlation coefficients between the modalities' representations. As shown by Setayesh et al. (2026), the optimal value of $f(\boldsymbol{z}_i^n, \{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M)$ is proportional to the density ratio $p(\boldsymbol{z}_i^n, \{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M)/p(\boldsymbol{z}_i^n)p(\{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M)$. We assume that the joint probability density of the concatenated representations of the $M$ modalities is a multivariate Gaussian distribution with mean $\mathbf{0} \in \mathbb{R}^{Md}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{P} \otimes \boldsymbol{I}_d$, where $\mathbf{P}$ is defined in (6). We empirically evaluate this assumption in Appendix B for settings with more than two modalities, using the Gaussianity diagnostics employed by Betser et al. (2026) for dual-modality settings.

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1M} \\ \rho_{12} & 1 & \cdots & \rho_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1M} & \rho_{2M} & \cdots & 1 \end{bmatrix}. \tag{6}$$

In Appendix C, we show that under the multivariate Gaussian distribution assumption, $p(\boldsymbol{z}_i^n, \{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M)/p(\boldsymbol{z}_i^n)p(\{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M) \propto \tilde{\alpha}_{\boldsymbol{j}} \exp(-\sum_{m\in[M]\setminus\{n\}} [\mathbf{P}^{-1}]_{n,m} \boldsymbol{z}_i^n \cdot \boldsymbol{z}_{j_m}^m)$ holds, where $\tilde{\alpha}_{\boldsymbol{j}} = \exp(-\frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\sum_{m'\in[M]\setminus\{n,m\}} \tilde{\gamma}_{m,m'}^{(M)} \boldsymbol{z}_{j_m}^m \cdot \boldsymbol{z}_{j_{m'}}^{m'})$, and $\tilde{\gamma}_{m,m'}^{(M)} = [\mathbf{P}^{-1}]_{m,m'} - [\mathbf{P}_{-n}^{-1}]_{m-\mathbb{1}_{\{m>n\}},\, m'-\mathbb{1}_{\{m'>n\}}}$. Here, $[\mathbf{P}^{-1}]_{n,m}$ denotes the $(n,m)$-th entry of $\mathbf{P}^{-1}$, $\mathbf{P}_{-n} \in \mathbb{R}^{(M-1)\times(M-1)}$ is the matrix obtained by removing the $n$-th row and column from $\mathbf{P}$, and $\mathbb{1}_{\{m>n\}}$ is the indicator function that equals one if $m > n$ and zero otherwise.

Based on the provided correlation-based formulation, we can observe that the temperature parameter $\tau_{nm}^{(M)}$ in the Muscle loss is related to the $(n,m)$-th entry of $\mathbf{P}^{-1}$ as follows: $\tau_{nm}^{(M)} = -1/[\mathbf{P}^{-1}]_{n,m}$. Additionally, the temperature parameter $\tau_{mm'}^{(M-1)}$ in the Muscle loss is related to an element of $\mathbf{P}_{-n}^{-1}$ as follows: $\tau_{mm'}^{(M-1)} = -1/[\mathbf{P}_{-n}^{-1}]_{m-\mathbb{1}_{\{m>n\}},\, m'-\mathbb{1}_{\{m'>n\}}}$. Accordingly, we define $\rho$-Muscle as follows:

$$\mathcal{L}_{\rho\text{-Muscle}}^n(\boldsymbol{z}_i^n) = -\log \frac{\tilde{\alpha}_{(i,\ldots,i)} \exp\left(-\boldsymbol{z}_i^n \cdot \sum_{m\in[M]\setminus\{n\}} [\mathbf{P}^{-1}]_{n,m} \boldsymbol{z}_i^m\right)}{\sum_{\boldsymbol{j}\in\mathcal{J}^n} \tilde{\alpha}_{\boldsymbol{j}} \exp\left(-\boldsymbol{z}_i^n \cdot \sum_{m\in[M]\setminus\{n\}} [\mathbf{P}^{-1}]_{n,m} \boldsymbol{z}_{j_m}^m\right)}. \tag{7}$$

Note that $\rho$-Muscle and $\rho$-InfoNCE are equivalent in the two-modality case. Thus, hereafter, we refer to $\rho$-Muscle as the correlation-based contrastive loss, as it applies to any number of modalities. In the next section, we present how $\rho$-Muscle facilitates multimodal representation learning by adaptively determining correlation coefficients between modality pairs as their representations are learned.

## 4 CorrCLR Methodology

In this section, we propose CorrCLR, a novel algorithm designed to learn multimodal representations while adaptively updating their correlation coefficients. The correlation-based contrastive loss formulations in Section 3 allow us to shift focus from tuning temperature parameters in CL to determining the correlation coefficients among multimodal representations. Unlike the complex dynamics of temperature parameters, correlation coefficients can be directly measured. Furthermore, they provide valuable insights into which modalities are tightly coupled and which are loosely related.

Algorithm 1 describes our proposed correlation-based CL of representations. CorrCLR consists of two main steps which are performed iteratively. In Step 1 (Lines 6–9), we update the correlation

**Algorithm 1** CorrCLR Algorithm

1: **Input:** Number of training epochs $E$; correlation coefficient update interval $T$; momentum parameter $\beta \in (0, 1)$; sampling ratio $r \in (0, 1]$; a positive definite initial correlation matrix $\mathbf{P}^0$.
2: **Output:** Learned parameters for each modality encoder; final correlation coefficient matrix $\mathbf{P}$.
3: Initialize $\mathbf{P} = \mathbf{P}^0$.
4: **for** each epoch $e \in [E]$ **do**
5: **if** $e \bmod T = 0$ **then** ▷ *Step 1*
6: Sample a subset of the training data with sampling ratio $r$ and compute the representations via modality encoders.
7: Perform feature-wise standardization on the representations of each modality.
8: Compute the empirical correlation matrix $\hat{\mathbf{P}}$ from the standardized modality representations.
9: $\mathbf{P} \leftarrow \beta\mathbf{P} + (1 - \beta)\hat{\mathbf{P}}$
10: **end if**
11: Update modality encoders by minimizing the proposed loss in (7), given $\mathbf{P}$. ▷ *Step 2*
12: **end for**

matrix $\mathbf{P}$, and in Step 2 (Line 11), we minimize the correlation-based contrastive loss (i.e., $\rho$-Muscle in (7)) given the updated $\mathbf{P}$. To ensure that the updated correlation matrix $\mathbf{P}$ in Step 1 is stable and adheres to the assumptions outlined in Section 3, we must execute a sequence of operations.

We first initialize $\mathbf{P}$ with a positive definite (PD) initial correlation matrix $\mathbf{P}^0$ (Line 3). Subsequently, $\mathbf{P}$ is updated every $T$ epochs, which serves as the update interval. During these $T$ epochs, the correlation-based contrastive loss is minimized using a fixed $\mathbf{P}$ to update the parameters of the modality encoders. This approach enhances the stability of the algorithm, especially when the encoders are trained from scratch. To update $\mathbf{P}$, we first compute $\hat{\mathbf{P}}$, the empirical correlation matrix derived from the modality representations (Line 8). Then, $\mathbf{P}$ is updated via a moving average with momentum $\beta$ (Line 9). This ensures a smooth transition for $\mathbf{P}$, promoting consistency during the training of the modality encoders and, as we discuss later, guaranteeing that $\mathbf{P}$ remains PD.

To compute the empirical correlation matrix $\hat{\mathbf{P}}$, we first sample a subset of the training data with a sampling ratio $r$ and obtain their representations using the updated modality encoders (Line 6). Let $\boldsymbol{Z}^m \in \mathbb{R}^{|\mathcal{D}_r| \times d}$ denote the representation matrix for modality $m \in [M]$ corresponding to the samples in the subset $\mathcal{D}_r \subseteq \mathcal{D}$, where $\mathcal{D}$ denotes the full training dataset. We then apply feature-wise standardization to the representation matrix $\boldsymbol{Z}^m$ of each modality $m$ (Line 7). This ensures that the representation vectors have zero mean and unit variance, which is consistent with the zero-mean Gaussian assumption and the correlation matrix with unit diagonal elements discussed in Section 3. Furthermore, feature-wise standardization prevents a specific feature or modality from becoming dominant when computing the empirical correlation matrix $\hat{\mathbf{P}}$ due to differences in feature scales or variances. Let $\tilde{\boldsymbol{Z}}^m$ denote the standardized representation matrix for modality $m$. The elements are computed as $[\tilde{\boldsymbol{Z}}^m]_{i,j} = ([\boldsymbol{Z}^m]_{i,j} - [\boldsymbol{\mu}^m]_j)/([\boldsymbol{\sigma}^m]_j + \epsilon)$, where $[\boldsymbol{\mu}^m]_j$ and $[\boldsymbol{\sigma}^m]_j$ denote the mean and standard deviation of the $j$-th feature across the representations of modality $m$, respectively. The term $\epsilon$ is a small constant added to the denominator to prevent division by zero. Finally, the empirical correlation coefficient $\hat{\rho}_{nm}$ within the matrix $\hat{\mathbf{P}}$ is computed as follows:

$$\hat{\rho}_{nm} = [\hat{\mathbf{P}}]_{n,m} = \frac{1}{|\mathcal{D}_r|d} \mathrm{Tr}\left[(\tilde{\boldsymbol{Z}}^n)^T \tilde{\boldsymbol{Z}}^m\right]. \tag{8}$$

In Appendix D, we show that $\hat{\mathbf{P}}$ is always a positive semi-definite (PSD) matrix; moreover, $\mathbf{P}$ is always PD. Consequently, CorrCLR guarantees that the inverse of $\mathbf{P}$ always exists and the correlation-based contrastive loss can be computed using the elements of $\mathbf{P}^{-1}$. We analyze the computational overhead of CorrCLR in Appendix E. In particular, we show that the update interval $T$ and sampling ratio $r$ provide flexibility in reducing the additional computational cost incurred by the alternating correlation matrix updates in CorrCLR.

## 5 Experiments

### 5.1 Experimental Setup

We conduct our experiments across three multimodal representation learning settings, including 3D point-image-text, video-audio-text, and mmWave-WiFi-RFID. We also report results for the multi-view setting, which involves multiple views from a single modality (i.e., images), in Appendix F.

**3D Point-Image-Text Setting** Our first multimodal setting involves three modalities: 3D point clouds (P), 2D images (I), and text descriptions (T). We adopt the OpenShape framework (Liu et al., 2023) to extract these modalities. ShapeNetCore (Chang et al., 2015) is used as training data, and zero-shot classification accuracy is evaluated on the ModelNet40 (Wu et al., 2015), ScanObjectNN (Uy et al., 2019), and Objaverse-LVIS (Deitke et al., 2023) datasets. Experiments are conducted using PointBERT (Yu et al., 2022) and SparseConv (Choy et al., 2019) as 3D encoders, while a pretrained OpenCLIP ViT-bigG-14 (Ilharco et al., 2021) serves as the image and text encoders. Following Liu et al. (2023) and Xue et al. (2023), the 3D encoders are trained from scratch, whereas the image and text encoders remain frozen during training. We use Top1, Top3, and Top5 zero-shot classification accuracies as performance metrics and compare CorrCLR with ULIP (Xue et al., 2023) and OpenShape (Liu et al., 2023), both of which employ pairwise alignment with a learnable temperature in the InfoNCE loss.

**Video-Audio-Text Setting** This setting consists of video frames (V), audio clips (A), and text captions (T). We employ the VAST framework (Chen et al., 2023b) to generate these modalities. A subset of 85K YouTube videos is extracted from the VAST-27M (Chen et al., 2023b) dataset for training, and zero-shot multimodal retrieval performance is evaluated on the MSR-VTT (Xu et al., 2016), DiDeMo (Anne Hendricks et al., 2017), and ActivityNet (Caba Heilbron et al., 2015) datasets. Following Chen et al. (2023b) and Cicchetti et al. (2025b), the V, A, and T encoders are initialized with pretrained EVA-CLIP-ViT-G (Sun et al., 2023), BEATs (Chen et al., 2023a), and BERT-Base (Devlin et al., 2019), respectively. Zero-shot recall at 1 and 10 (R@1 and R@10) is used for all text-video (TV and VT), text-audio (TA and AT), and audio-video (AV and VA) retrieval tasks as the performance metric. Inspired by LanguageBind (Zhu et al., 2024), pairwise alignment across all modality pairs is considered as a baseline in this setting.

**mmWave-WiFi-RFID Setting** Our third setting comprises three radio frequency (RF) modalities: mmWave (M), WiFi (W), and RFID (R). We utilize the XRF55 dataset (Wang et al., 2024) to evaluate multimodal representations for human action recognition in both intra-domain and cross-domain scenarios. The RF signals are collected from four indoor scenes. For the intra-domain scenario, we follow the standard train/test split of the XRF55 dataset, while for the cross-domain scenario, the first three scenes are used for training and the fourth for testing. Following Chen & Yang (2025), each modality encoder is initialized with a corresponding ResNet-18 (He et al., 2016) pretrained on the training dataset. We perform full fine-tuning and measure human action classification accuracy as the performance metric. Additionally, the recently proposed Muscle loss (Setayesh et al., 2026) with fixed temperatures and feature fusion (Chen & Yang, 2025) are considered as baselines.

**Implementation Details** In all three multimodal representation learning settings, we also compare the performance of CorrCLR with that of iSogCLR (Qiu et al., 2023), Symile (Saporta et al., 2024), GRAM (Cicchetti et al., 2025b), and TRIANGLE (Cicchetti et al., 2025a) as additional baselines. We implement CorrCLR and the baseline algorithms using PyTorch (Paszke et al., 2019). Unless otherwise stated, we set $r$ and $\beta$ to 1 and 0.9, respectively, in Algorithm 1. We also set $T$ to 10, 1, and 5 for the 3D point-image-text, video-audio-text, and mmWave-WiFi-RFID, respectively. For all settings, we provide an overall evaluation metric, denoted as $\Delta$, which facilitates the comparison of different methods across heterogeneous tasks relative to a base method (Lu et al., 2024; Setayesh et al., 2026). Specifically, we define $\Delta = \frac{1}{N}\sum_{n\in[N]}(\eta^n_{\text{method}} - \eta^n_{\text{base}})/\eta^n_{\text{base}}$, where $\eta^n_{\text{method}}$ and $\eta^n_{\text{base}}$ represent the performance metrics of the evaluated method and the base method, respectively, for task $n \in [N]$. All metrics are reported in percentages (%), with higher values being preferable. Additional implementation details are provided in Appendix G.

### 5.2 Benchmark Experiments

**Comparison with Baselines in the 3D Point-Image-Text Setting** Table 1 shows that CorrCLR provides the best overall performance in terms of $\Delta$ among all baselines. In particular, CorrCLR achieves zero-shot classification accuracies that are higher than or comparable to those of strong baselines such as iSogCLR and GRAM across different test datasets and 3D encoders. Furthermore, Fig. 2 illustrates how correlation coefficients among multimodal representations evolve during CorrCLR training. This figure indicates that the correlation coefficient between the P and I modalities is higher than those of the other pairs, which aligns with the expectation that 3D point clouds may be more correlated with 2D images than with text descriptions. Additionally, because correlation coefficients are inherent characteristics of the data, their values do not change significantly when the 3D encoder is changed. In Appendix H, using synthetically generated data, we further show that

Table 1: Comparison of CorrCLR with baseline methods in the 3D point-image-text setting. Symile is omitted from the table because its performance is substantially lower than that of the other methods in this setting.

| 3D Encoder | Method | ModelNet40 | | | ScanObjectNN | | | Objaverse-LVIS | | | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Top1 | Top3 | Top5 | Top1 | Top3 | Top5 | Top1 | Top3 | Top5 | |
| PointBERT | ULIP (Xue et al., 2023) | 60.4 | 79.0 | 84.4 | 51.5 | 71.1 | 80.2 | 6.2 | 13.6 | 17.9 | 0.00 |
| | OpenShape (Liu et al., 2023) | 64.8 | 80.7 | 85.7 | 46.6 | 67.8 | 78.3 | 8.7 | 16.8 | 21.0 | +8.40 |
| | iSogCLR (Qiu et al., 2023) | 69.2 | 85.0 | 90.4 | 51.5 | 72.9 | 83.1 | 11.5 | 20.6 | 25.2 | +23.68 |
| | GRAM (Cicchetti et al., 2025b) | 69.5 | 83.6 | 88.4 | 55.4 | 75.2 | 87.8 | 11.8 | 19.8 | 23.9 | +24.21 |
| | TRIANGLE (Cicchetti et al., 2025a) | 65.0 | 81.0 | 86.8 | 46.8 | 70.3 | 81.1 | 10.2 | 17.9 | 22.0 | +13.66 |
| | CorrCLR (Ours) | 70.6 | 87.8 | 91.7 | 51.0 | 74.7 | 86.3 | 11.9 | 20.0 | 24.3 | **+24.79** |
| SparseConv | OpenShape (Liu et al., 2023) | 70.4 | 85.0 | 88.9 | 53.4 | 75.1 | 84.3 | 10.0 | 19.9 | 24.9 | +21.18 |
| | iSogCLR (Qiu et al., 2023) | 71.4 | 88.1 | 93.0 | 56.4 | 73.7 | 84.7 | 12.6 | 22.8 | 28.2 | +31.90 |
| | GRAM (Cicchetti et al., 2025b) | 73.5 | 87.2 | 91.5 | 60.4 | 80.3 | 90.8 | 12.0 | 21.1 | 25.9 | +30.81 |
| | TRIANGLE (Cicchetti et al., 2025a) | 69.4 | 86.5 | 90.6 | 57.5 | 75.8 | 87.8 | 11.6 | 20.7 | 25.8 | +26.99 |
| | CorrCLR (Ours) | 71.7 | 88.1 | 92.5 | 60.1 | 80.4 | 91.2 | 12.8 | 21.6 | 26.7 | **+33.08** |

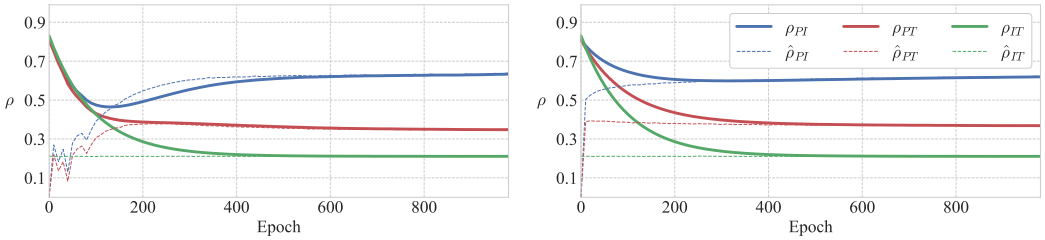



Figure 2: Evolution of updated and empirical correlation coefficients between representations of the P, I, and T modalities in the 3D point-image-text setting, using PointBERT (left) and SparseConv (right).

the correlation coefficients among multimodal representations estimated by CorrCLR closely match those among the original multimodal data samples.

**Comparison with Baselines in the Video-Audio-Text Setting** Table 2 presents the improvements achieved by CorrCLR over the baseline methods on zero-shot retrieval tasks using the MSR-VTT dataset. Results for the DiDeMo and ActivityNet datasets are provided in Appendix I. The superior overall performance of CorrCLR in terms of $\Delta$ indicates its effectiveness in capturing dependencies among multimodal representations. This improvement stems from the use of $\rho$-Muscle and the adaptive estimation of correlation coefficients that align well with the underlying data distributions. We present the evolution of the correlation coefficients for this setting in Appendix J.1.

Table 2: Comparison of CorrCLR with baseline methods in the video-audio-text setting on MSR-VTT.

| Dataset | Method | TV | | VT | | TA | | AT | | AV | | VA | | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | |
| MSR-VTT | Pairwise Alignment | 16.7 | 53.2 | 14.2 | 50.2 | 2.0 | 14.4 | 1.8 | 13.1 | 5.8 | 29.6 | 6.3 | 28.2 | 0.00 |
| | iSogCLR (Qiu et al., 2023) | 16.9 | 54.3 | 19.5 | 52.2 | 4.1 | 17.3 | 2.2 | 15.5 | 7.9 | 31.3 | 6.5 | 29.2 | +21.58 |
| | Symile (Saporta et al., 2024) | 17.8 | 55.4 | 19.1 | 51.2 | 3.7 | 16.3 | 1.8 | 12.2 | 7.9 | 31.3 | 7.1 | 34.2 | +17.87 |
| | GRAM (Cicchetti et al., 2025b) | 15.2 | 50.1 | 13.2 | 47.7 | 3.9 | 16.5 | 3.7 | 17.7 | 11.0 | 42.4 | 13.6 | 41.7 | +43.34 |
| | TRIANGLE (Cicchetti et al., 2025a) | 37.3 | 75.1 | 37.4 | 73.6 | 3.5 | 14.8 | 1.5 | 10.4 | 4.2 | 26.2 | 2.3 | 27.4 | +25.80 |
| | CorrCLR (Ours) | 23.6 | 60.5 | 21.5 | 57.6 | 4.3 | 20.9 | 3.8 | 18.6 | 10.3 | 35.6 | 8.3 | 36.0 | **+49.31** |

**Comparison with Baselines in the mmWave-WiFi-RFID Setting** Table 3 shows the classification accuracy for human action recognition in both intra-domain and cross-domain scenarios across various modality combinations. The feature fusion method (Chen & Yang, 2025) simply concatenates the representations from the available modalities and fuses them using an MLP. In contrast, for contrastive methods, we first align the representations from the M, W, and R encoders through CL before performing the subsequent fusion. Based on the results in Table 3, our observations are as follows: (1) CorrCLR outperforms the baseline methods in terms of $\Delta$. (2) Representation alignment prior to fusion significantly improves performance. (3) While the Muscle loss requires six temperature parameters tuned via grid search, CorrCLR eliminates this requirement by adaptively estimating the correlation coefficients. (4) RFID operates at a lower frequency. Thus, its backscatter measurements are relatively coarse and lack the resolution needed to capture complex human actions effectively (Zhang et al., 2025). The evolution of correlation coefficients during CorrCLR training (see Appendix J.2) further indicates that $\rho_{\text{MR}}$ and $\rho_{\text{WR}}$ are lower than $\rho_{\text{MW}}$.

### 5.3 Ablation Studies

We empirically demonstrate that although the CorrCLR algorithm involves four hyperparameters (i.e., $\mathbf{P}^0$, $\beta$, $r$, and $T$), its performance is relatively insensitive to the choices of $\mathbf{P}^0$, $\beta$, and $r$. Additionally, choosing a small to moderate value for $T$ (e.g., between 1 and 10) is sufficient to

Table 3: Comparison of CorrCLR with baseline methods in the mmWave-WiFi-RFID setting.

| Scenario | Method | Modality | | | | | | | Δ |
|---|---|---|---|---|---|---|---|---|---|
| | | M | W | R | M + W | M + R | W + R | M + W + R | |
| Intra-Domain | Feature Fusion (Chen & Yang, 2025) | 87.1 | 76.5 | 51.8 | 85.5 | 54.1 | 55.4 | 57.9 | 0.00 |
| | iSogCLR (Qiu et al., 2023) | 87.8 | 86.1 | 55.7 | 87.5 | 88.0 | 84.6 | 84.1 | +26.24 |
| | Symile (Saporta et al., 2024) | 87.8 | 81.7 | 54.5 | 85.3 | 54.6 | 72.2 | 75.5 | +10.58 |
| | GRAM (Cicchetti et al., 2025b) | 86.1 | 82.5 | 58.5 | 84.8 | 83.6 | 83.8 | 90.5 | +25.79 |
| | TRIANGLE (Cicchetti et al., 2025a) | 87.0 | 84.7 | 55.1 | 86.7 | 87.3 | 84.4 | 84.6 | +25.44 |
| | Muscle (Setayesh et al., 2026) | 88.1 | 81.8 | 51.6 | 87.4 | 87.2 | 78.9 | 81.5 | +22.00 |
| | CorrCLR (Ours) | 87.0 | 84.5 | 58.1 | 91.6 | 87.1 | 83.9 | 90.3 | **+28.26** |
| Cross-Domain | Feature Fusion (Chen & Yang, 2025) | 34.7 | 2.7 | 1.7 | 2.7 | 1.5 | 1.7 | 1.9 | 0.00 |
| | iSogCLR (Qiu et al., 2023) | 30.9 | 3.5 | 2.2 | 2.6 | 27.1 | 3.8 | 2.4 | +267.82 |
| | Symile (Saporta et al., 2024) | 30.7 | 2.5 | 1.6 | 3.7 | 19.7 | 2.4 | 3.7 | +191.73 |
| | GRAM (Cicchetti et al., 2025b) | 29.2 | 4.4 | 2.3 | 28.1 | 27.3 | 2.6 | 8.2 | +442.71 |
| | TRIANGLE (Cicchetti et al., 2025a) | 28.9 | 2.2 | 2.7 | 2.2 | 30.1 | 2.6 | 4.2 | +293.79 |
| | Muscle (Setayesh et al., 2026) | 31.0 | 2.7 | 2.4 | 15.7 | 32.3 | 2.9 | 3.3 | +381.91 |
| | CorrCLR (Ours) | 22.2 | 5.3 | 2.3 | 29.5 | 28.9 | 6.9 | 23.3 | **+615.23** |

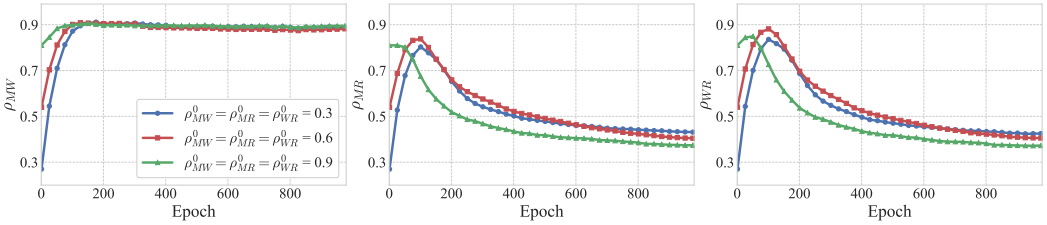



Figure 3: Evolution of $\rho_{\text{MW}}$, $\rho_{\text{MR}}$, and $\rho_{\text{WR}}$ for different initial values. CorrCLR is robust to the choice of initial correlation coefficients.

ensure good convergence and high performance. All ablation studies are conducted for the intra-domain scenario within the mmWave-WiFi-RFID setting. The ablation studies for $\beta$ and $T$ are provided in Appendices K.2 and K.4, respectively.

**Effect of the Initial Correlation Matrix $\mathbf{P}^0$** Fig. 3 illustrates the final correlation coefficients determined by CorrCLR for each pair of modalities across three different initializations of $\mathbf{P}^0$. The results demonstrate that CorrCLR is robust to the initial correlation matrix. Specifically, CorrCLR converges to nearly identical values for $\rho_{\text{MW}}$, $\rho_{\text{MR}}$, and $\rho_{\text{WR}}$ regardless of the initialization. Furthermore, as shown in Appendix K.1, the overall performance of CorrCLR in human action recognition is not significantly affected by different correlation matrix initializations.

**Effect of the Sampling Ratio $r$** We sample a subset of the training data using a sampling ratio $r$ to compute the empirical correlation matrix $\hat{\mathbf{P}}$. A lower $r$ indicates that fewer samples are used to compute this matrix. Since the empirical correlation matrix reflects a statistical property of the data, a lower $r$ is still sufficient to provide an accurate estimate of $\hat{\mathbf{P}}$. In Appendix K.3, we show that the correlation coefficients obtained by CorrCLR for each pair of modalities remain relatively stable across different choices of $r$. Furthermore, the results in Table 4 demonstrate that the performance of CorrCLR in human action recognition is highly robust to the choice of $r$.

Table 4: Performance of CorrCLR for different $r$ values.

| $r$ | Modality | | | | | | | Δ |
|---|---|---|---|---|---|---|---|---|
| | M | W | R | M + W | M + R | W + R | M + W + R | |
| 0.2 | 86.3 | 84.8 | 57.9 | 91.7 | 86.9 | 84.7 | 90.9 | +28.47 |
| 0.4 | 87.5 | 84.3 | 56.9 | 87.6 | 87.4 | 85.1 | 88.0 | +27.11 |
| 0.6 | 86.4 | 84.2 | 54.9 | 89.1 | 86.9 | 84.6 | 90.6 | +27.01 |
| 0.8 | 86.5 | 85.0 | 58.7 | 90.6 | 87.1 | 85.6 | 91.3 | +28.96 |
| 1 | 87.0 | 84.5 | 58.1 | 91.6 | 87.1 | 83.9 | 90.3 | +28.26 |

## 6 Conclusion

In this work, we addressed a fundamental question: how the heuristic temperature hyperparameter in contrastive loss functions, such as InfoNCE, can be replaced with an interpretable parameter to facilitate multimodal representation learning. We demonstrated that the temperature hyperparameters in the InfoNCE and Muscle loss functions are inherently related to the correlation coefficients of multimodal representations. Based on these insights, we proposed $\rho$-InfoNCE and $\rho$-Muscle as correlation-based counterparts to the InfoNCE and Muscle loss functions, respectively. Building on these reformulations, we introduced CorrCLR, a framework that learns modality representations while adaptively updating their correlation coefficients as measurable parameters. Through extensive experiments, we showed that CorrCLR provides robust performance improvements over state-of-the-art baselines across diverse multimodal settings and downstream tasks. Discussions of future work and limitations are provided in Appendices M and N, respectively.

### AI USE STATEMENT

In this work, we used generative AI tools to edit our paper to improve readability, identify relevant literature, and edit our code. We did not use generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, or implement methods. Additionally, cleaning and reformatting datasets and supporting qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

For theoretical results, we provide the complete proofs along with their underlying assumptions in Section 3 and Appendices A, C, and D. All data used in this study are publicly available. The code for running our algorithm is also provided in the supplementary material. Furthermore, we provide all experimental details in Section 5.1 and Appendices F and G. We have cited all original sources for the models, baselines, and datasets used in this work. Please refer to Section 5.1 for further details. While resource constraints preclude exhaustive multi-seed evaluations across all experiments, we report multi-seed results for a representative benchmark setting in Appendix L. The consistency of performance gaps across diverse datasets and settings strongly suggests that the observed improvements reflect genuine gains rather than stochastic variation.

## REFERENCES

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017.

Parikshit Bansal, Ali Kavis, et al. Understanding contrastive learning via Gaussian mixture models. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, San Diego, CA, Dec. 2025.

Roy Betser, Eyal Gofer, Meir Yossef Levi, and Guy Gilboa. InfoNCE induces Gaussian distribution. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Rio de Janeiro, Brazil, Apr. 2026.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, Jun. 2015.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, Dec. 2015.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, Jul. 2023a.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, Dec. 2023b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, Mar. 2020b.

Xinyan Chen and Jianfei Yang. X-Fi: A modality-invariant foundation model for multimodal human sensing. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Singapore, Apr. 2025.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, Jun. 2019.

Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal alignment beyond cosine similarity. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, San Diego, CA, Dec. 2025a.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Singapore, Apr. 2025b.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, Jun. 2023.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22, 1977.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. North Amer. chapter Assoc. Comput. Linguistics: Human Lang. Technologies (NAACL-HLT)*, Minneapolis, MN, Jun 2019.

Persi Diaconis and David Freedman. A dozen de Finetti-style results in search of a theory. In *Annales de l'IHP Probabilités et statistiques*, volume 23, pp. 397–423, 1987.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Singapore, Apr. 2025.

Xiyuan Gao, Bing Cao, Pengfei Zhu, Nannan Wang, and Qinghua Hu. Asymmetric reinforcing against multi-modal representation bias. In *Proc. AAAI Conf. Artif. Intell.*, Philadelphia, PA, Feb. 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, Jun. 2023.

Hadi Hadizadeh, S Faegheh Yeganli, Bahador Rashidi, and Ivan V Bajić. Mutual information analysis in multimodal learning systems. In *Proc. IEEE Int. Conf. Multimedia Information Processing and Retrieval (MIPR)*, San Jose, CA, Aug. 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, Jun. 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021.

Zizheng Huang, Haoxing Chen, Ziqi Wen, Chao Zhang, Huaxiong Li, Bo Wang, and Chunlin Chen. Model-aware contrastive learning: Towards escaping the dilemmas. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, Jul. 2023.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, jul. 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021.

Bum Jun Kim and Sang Woo Kim. Temperature-free loss function for contrastive learning. *arXiv preprint arXiv:2501.17683*, Jan. 2025.

Michael Kirchhof, Enkelejda Kasneci, and Seong Joon Oh. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, Jul. 2023.

David Klindt, Yann LeCun, and Randall Balestriero. When does LeJEPA learn a world model? *arXiv preprint arXiv:2605.26379*, May 2026.

Panagiotis Koromilas, Efthymios Georgiou, Giorgos Bouritsas, Theodoros Giannakopoulos, Mihalis A Nicolaou, and Yannis Panagakis. A principled framework for multi-view contrastive learning. *arXiv preprint arXiv:2507.06979*, Jul. 2025.

Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021.

Yangmin Li, Ruiqi Zhu, and Wengen Li. CorMulT: A semi-supervised modality correlation-aware multimodal transformer for sentiment analysis. *IEEE Trans. Affect. Comput.*, 16(3):2321–2333, Jul. 2025.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3d shape representation towards open-world understanding. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, Dec. 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, May 2019.

Yiwei Lu, Guojun Zhang, Sun Sun, Hongyu Guo, and Yaoliang Yu. $f$-MICL: Understanding and generalizing InfoNCE-based contrastive learning. *Trans. Mach. Learn. Res.*, Oct. 2023.

Yuxiang Lu, Suizhi Huang, Yuwen Yang, Shalayiding Sirejiding, Yue Ding, and Hongtao Lu. FedHCA$^2$: Towards hetero-client federated multi-task learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle WA, Jun. 2024.

Siladittya Manna, Soumitri Chattopadhyay, Rakesh Dey, Umapada Pal, and Saumik Bhattacharya. Dynamically scaled temperature in self-supervised contrastive learning. *IEEE Trans. Artif. Intell.*, 6(6):1502–1512, Jun. 2025.

Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. Drop-upcycling: Training sparse mixture of experts with partial re-initialization. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Singapore, Apr. 2025.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Bellevue, WA, Jun. 2011.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, Jan. 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, Jul. 2023.

Zi-Hao Qiu, Siqi Guo, Mao Xu, Tuo Zhao, Lijun Zhang, and Tianbao Yang. To cool or not to cool? Temperature network meets large foundation models via DRO. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, Jul. 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021.

Md Aamir Raihan and Tor Aamodt. Sparse weight activation training. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2020.

Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, Feb. 2023.

Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, Sipeng Zheng, and Qin Jin. Accommodating audio modality in CLIP for multimodal processing. In *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, Feb. 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115:211–252, Apr. 2015.

Adriel Saporta, Aahlad Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with Symile: Simple model-agnostic representation learning for unlimited modalities. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2024.

Mehdi Setayesh, Mahdi Beitollahi, Yasser H. Khalil, and Hongliang Li. Toward enhancing representation learning in federated multi-task settings. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Rio de Janeiro, Brazil, Apr. 2026.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, et al. Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Singapore, Apr. 2025.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, Mar. 2023.

Igor Susmelj, Matthias Heller, Philipp Wirth, Jeremy Prescott, Malte Ebner, and et al. Lightly, n.d. URL `https://github.com/lightly-ai/lightly`.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Vancouver, Canada, Jul. 2025.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea, Oct. 2019.

Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. XRF55: A radio frequency dataset for human indoor action analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1): 1–34, Mar. 2024.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, Jun. 2021.

CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, 11(1): 95–103, Mar. 1983.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, Jun. 2015.

Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. In *Proc. Advances in Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, Dec. 2023.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, Jun. 2016.

Qinfu Xu, Yiwei Wei, Chunlei Wu, Leiquan Wang, Shaozu Yuan, Jie Wu, Jing Lu, and Hengyang Zhou. Towards multimodal sentiment analysis via hierarchical correlation modeling with semantic distribution constraints. In *Proc. AAAI Conf. Artif. Intell.*, Philadelphia, PA, Feb. 2025.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, Jun. 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. ULIP-2: Towards scalable multimodal pre-training for 3d understanding. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle WA, Jun. 2024.

Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. i-Code: An integrative and composable multimodal learning framework. In *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, Feb. 2023.

Wenzhe Yin, Pan Zhou, Zehao Xiao, Jie Liu, Shujian Yu, Jan-Jakob Sonke, and Efstratios Gavves. Towards uniformity and alignment for multimodal representation learning. *arXiv preprint arXiv:2602.09507*, Feb. 2026.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3d point cloud transformers with masked point modeling. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, Jun. 2022.

Zhuoning Yuan, Dixian Zhu, Zi-Hao Qiu, Gang Li, Xuanhui Wang, and Tianbao Yang. LibAUC: A deep learning library for X-risk optimization. In *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Long Beach, CA, Aug. 2023.

Jie Zhang, Zuan Qin, Bingxun Mu, Lang He, Zhongmin Wang, and Jing Chen. Robust cross-domain RF-based multimodal activity recognition with few-shot adaptation. *IEEE Internet Things J.*, 12 (24):54899–54920, Dec. 2025.

Yiyun Zhou, Mingjing Xu, Jingwei Shi, Quanjiang Li, and Jingyuan Chen. Collaborative representation learning for alignment of tactile, language, and vision modalities. In *Proc. AAAI Conf. Artif. Intell.*, Singapore, Jan. 2026.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2024.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. CMAD: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Honolulu, HI, Oct. 2025.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021.

We provide more details and results about our work in the appendices. Here are the contents:

- Appendix A: Density ratio for the alignment of two modalities.
- Appendix B: Experimental evidence for the Gaussian and covariance structure assumptions.
- Appendix C: Density ratio for the alignment of multiple modalities.
- Appendix D: Invertibility of matrix $\mathbf{P}$.
- Appendix E: Computational cost analysis of CorrCLR.
- Appendix F: Experimental results for the multiview setting.
- Appendix G: Additional implementation details.
- Appendix H: Correlation coefficients estimated by CorrCLR using synthetic data.
- Appendix I: Additional experimental results for the video-audio-text setting.
- Appendix J: Evolution of correlation coefficients during CorrCLR training.
- Appendix K: Extended ablation studies.
- Appendix L: Multiple-seed tests for the intra-domain scenario in the mmWave-WiFi-RFID setting.
- Appendix M: Discussion on possible future work directions.
- Appendix N: Limitations.

## A Density Ratio for the Alignment of Two Modalities

The multivariate Gaussian distribution of the random vector $\boldsymbol{z} \in \mathbb{R}^{2d}$ is defined by the following density function (Bishop, 2006):

$$p(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^{2d}|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right), \tag{9}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{2d}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{2d\times 2d}$ are the mean vector and the covariance matrix, respectively. Given the assumptions mentioned in Section 3.1, we have $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_i^n \\ \boldsymbol{z}_j^m \end{bmatrix}$, $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{nm} \\ \rho_{nm} & 1 \end{bmatrix} \otimes \boldsymbol{I}_d$. Thus, $p(\boldsymbol{z}_i^n, \boldsymbol{z}_j^m)$ is obtained as follows:

$$p(\boldsymbol{z}_i^n, \boldsymbol{z}_j^m) = \frac{1}{\sqrt{(2\pi)^{2d}(1-\rho_{nm}^2)^d}} \exp\left(-\frac{1}{2(1-\rho_{nm}^2)}\left(\|\boldsymbol{z}_i^n\|^2 - 2\rho_{nm}\boldsymbol{z}_i^n\cdot\boldsymbol{z}_j^m + \|\boldsymbol{z}_j^m\|^2\right)\right). \tag{10}$$

Since the multivariate Gaussian distribution is closed under marginalization (Bishop, 2006), we have $p(\boldsymbol{z}_i^n) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}\|\boldsymbol{z}_i^n\|^2\right)$ and $p(\boldsymbol{z}_j^m) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}\|\boldsymbol{z}_j^m\|^2\right)$. Given that in CL, the representation vectors $\boldsymbol{z}_i^n$ and $\boldsymbol{z}_j^m$ are normalized to unit norm before calculating their cosine similarities, we have

$$\frac{p(\boldsymbol{z}_i^n, \boldsymbol{z}_j^m)}{p(\boldsymbol{z}_i^n)p(\boldsymbol{z}_j^m)} \propto \exp\left(\frac{\rho_{nm}}{1-\rho_{nm}^2}\boldsymbol{z}_i^n\cdot\boldsymbol{z}_j^m\right). \tag{11}$$

## B Experimental Evidence for the Gaussian and Covariance Structure Assumptions

To evaluate the Gaussianity assumption, we use complementary diagnostics that target the radial and coordinate-wise behavior of multimodal representations. In particular, we use the Gaussianity diagnostics employed by Betser et al. (2026). These diagnostics include the coefficient of variation (CV), the Anderson–Darling (AD) test, and the D'Agostino–Pearson (DP) test. The CV measures norm concentration, with a lower CV indicating a distribution closer to Gaussian. For the AD and DP tests, we report the average test statistic and the fraction of compliant coordinates.

Lower AD values, higher DP values, and a higher fraction of compliant coordinates indicate better coordinate-wise Gaussianity. We conduct these experiments on representations obtained from pretrained models trained using CorrCLR and InfoNCE across 3D point-image-text and multiview settings described in Section 5.1 and Appendix F, respectively. Furthermore, we evaluate models trained on synthetically generated data derived from binary and Gaussian mixture model (GMM) distributions. The results in Table 5 show that the representations learned by CL in settings with more than two modalities also follow a Gaussian distribution.

Table 5: Gaussianity diagnostics of representations obtained from pretrained models.

| Metric | Synthetic Binary Data (3 Views) with CorrCLR | Synthetic GMM Data (3 Views) with CorrCLR | 3D Point-Image-Text with CorrCLR | 3D Point-Image-Text with Pairwise Alignment | Multiview (2 views) with CorrCLR | Multiview (2 views) with InfoNCE | Multiview (2 views) with Supervised Fine-Tuning |
|---|---|---|---|---|---|---|---|
| CV | 0.063 | 0.057 | 0.042 | 0.041 | 0.191 | 0.165 | 0.272 |
| AD Avg. ($< 0.752$) | 0.380 | 0.382 | 0.461 | 0.436 | 0.462 | 0.413 | 1.980 |
| AD Norm. Feat. (%) | 95.1 | 95.0 | 87.8 | 90.2 | 91.9 | 92.1 | 32.9 |
| DP Avg. ($> 0.05$) | 0.508 | 0.505 | 0.436 | 0.455 | 0.450 | 0.461 | 0.110 |
| DP Norm. Feat. (%) | 94.0 | 94.0 | 88.9 | 90.3 | 89.2 | 88.3 | 28.6 |
| Gaussian? | Yes | Yes | Yes | Yes | Yes | Yes | No |

To provide empirical evidence for the covariance structure assumption considered in Section 3.2, we compute descriptive statistics, including the mean, minimum, maximum, and standard deviation of the correlation coefficients between representations of modality pairs across all dimensions. Table 6 shows that the standard deviation falls within a range such that assuming the same correlation coefficient across all dimensions appears reasonable. Note that the considered covariance structure is a practical assumption that performs well across all our experiments. Furthermore, it keeps the number of correlation coefficients in $\rho$-Muscle on the same order as the number of temperature parameters used in the Muscle loss function and enhances interpretability by assigning a single correlation coefficient to each pair of modalities.

Table 6: Descriptive statistics of correlation variability across representation dimensions in 3D point-image-text setting.

| Modality Pairs | Mean | Minimum | Maximum | Standard Deviation |
|---|---|---|---|---|
| (3D Point, Image) | 0.622 | 0.335 | 0.844 | 0.065 |
| (3D Point, Text) | 0.367 | -0.023 | 0.675 | 0.098 |
| (Image, Text) | 0.211 | -0.117 | 0.521 | 0.086 |

## C Density Ratio for the Alignment of Multiple Modalities

Let $\boldsymbol{z} = \left[(\boldsymbol{z}_{j_1}^1)^T \quad (\boldsymbol{z}_{j_2}^2)^T \quad \dots \quad (\boldsymbol{z}_i^n)^T \quad \dots \quad (\boldsymbol{z}_{j_M}^M)^T\right]^T$. Given the assumptions mentioned in Section 3.2, $p\left(\boldsymbol{z}_i^n,\, \{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M\right)$ is obtained as follows:

$$
\begin{aligned}
&p\left(\boldsymbol{z}_i^n,\, \{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M\right) \\
&= \frac{1}{\sqrt{(2\pi)^{Md}|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right) \\
&= \frac{1}{\sqrt{(2\pi)^{Md}|\mathbf{P}|^d}} \exp\left(-\frac{1}{2}\left[\mathbf{P}^{-1}\right]_{n,n}\|\boldsymbol{z}_i^n\|^2 - \frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\left[\mathbf{P}^{-1}\right]_{m,m}\|\boldsymbol{z}_{j_m}^m\|^2\right) \\
&\times \exp\left(-\sum_{m\in[M]\setminus\{n\}}\left[\mathbf{P}^{-1}\right]_{n,m}\boldsymbol{z}_i^n\cdot\boldsymbol{z}_{j_m}^m - \frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\sum_{m'\in[M]\setminus\{n,m\}}\left[\mathbf{P}^{-1}\right]_{m,m'}\boldsymbol{z}_{j_m}^m\cdot\boldsymbol{z}_{j_{m'}}^{m'}\right),
\end{aligned}
\tag{12}
$$

where $\left[\mathbf{P}^{-1}\right]_{n,m}$ denotes the element in the $n$-th row and $m$-th column of the inverse matrix $\mathbf{P}^{-1}$.

Since the multivariate Gaussian distribution is closed under marginalization (Bishop, 2006), the joint density of the concatenated representations for the $M-1$ non-anchor modalities is also a multivariate Gaussian distribution with mean $\mathbf{0}\in\mathbb{R}^{(M-1)d}$ and covariance matrix $\mathbf{P}_{-n}\otimes\boldsymbol{I}_d$. Here,

$\mathbf{P}_{-n} \in \mathbb{R}^{(M-1)\times(M-1)}$ is the matrix obtained by removing the $n$-th row and $n$-th column from $\mathbf{P}$. We have

$$
\begin{aligned}
& p\left(\{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M\right) \\
&= \frac{1}{\sqrt{(2\pi)^{(M-1)d}|\mathbf{P}_{-n}|^d}} \exp\left(-\frac{1}{2}\sum_{m\in[M]\setminus\{n\}} \left[\mathbf{P}_{-n}^{-1}\right]_{m-\mathbb{1}_{\{m>n\}},m-\mathbb{1}_{\{m>n\}}} \|\boldsymbol{z}_{j_m}^m\|^2\right) \\
&\qquad \times \exp\left(-\frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\sum_{m'\in[M]\setminus\{n,m\}} \left[\mathbf{P}_{-n}^{-1}\right]_{m-\mathbb{1}_{\{m>n\}},m'-\mathbb{1}_{\{m'>n\}}} \boldsymbol{z}_{j_m}^m \cdot \boldsymbol{z}_{j_{m'}}^{m'}\right),
\end{aligned} \tag{13}
$$

where $\mathbb{1}_{\{m>n\}}$ denotes an indicator function that equals one if $m > n$ and zero otherwise.

Additionally, we have $p(\boldsymbol{z}_i^n) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}\|\boldsymbol{z}_i^n\|^2\right)$. Since representation vectors in CL are normalized to unit norm prior to calculating cosine similarities, the density ratio is derived using (12) and (13) as follows:

$$
\begin{aligned}
& \frac{p(\boldsymbol{z}_i^n,\, \{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M)}{p(\boldsymbol{z}_i^n)p(\{\boldsymbol{z}_{j_m}^m\}_{m=1,\, m\neq n}^M)} \\
&\propto \frac{\exp\left(-\sum_{m\in[M]\setminus\{n\}}\left[\mathbf{P}^{-1}\right]_{n,m}\boldsymbol{z}_i^n\cdot\boldsymbol{z}_{j_m}^m - \frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\sum_{m'\in[M]\setminus\{n,m\}}\left[\mathbf{P}^{-1}\right]_{m,m'}\boldsymbol{z}_{j_m}^m\cdot\boldsymbol{z}_{j_{m'}}^{m'}\right)}{\exp\left(-\frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\sum_{m'\in[M]\setminus\{n,m\}}\left[\mathbf{P}_{-n}^{-1}\right]_{m-\mathbb{1}_{\{m>n\}},m'-\mathbb{1}_{\{m'>n\}}}\boldsymbol{z}_{j_m}^m\cdot\boldsymbol{z}_{j_{m'}}^{m'}\right)} \\
&= \frac{\exp\left(-\sum_{m\in[M]\setminus\{n\}}\left[\mathbf{P}^{-1}\right]_{n,m}\boldsymbol{z}_i^n\cdot\boldsymbol{z}_{j_m}^m\right)}{\exp\left(-\frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\sum_{m'\in[M]\setminus\{n,m\}}\left(\left[\mathbf{P}_{-n}^{-1}\right]_{m-\mathbb{1}_{\{m>n\}},m'-\mathbb{1}_{\{m'>n\}}} - \left[\mathbf{P}^{-1}\right]_{m,m'}\right)\boldsymbol{z}_{j_m}^m\cdot\boldsymbol{z}_{j_{m'}}^{m'}\right)} \\
&= \tilde{\alpha}_{\boldsymbol{j}} \exp\left(-\sum_{m\in[M]\setminus\{n\}}\left[\mathbf{P}^{-1}\right]_{n,m}\boldsymbol{z}_i^n\cdot\boldsymbol{z}_{j_m}^m\right),
\end{aligned} \tag{14}
$$

where $\tilde{\alpha}_{\boldsymbol{j}} = \exp\left(-\frac{1}{2}\sum_{m\in[M]\setminus\{n\}}\sum_{m'\in[M]\setminus\{n,m\}}\tilde{\gamma}_{m,m'}^{(M)}\boldsymbol{z}_{j_m}^m\cdot\boldsymbol{z}_{j_{m'}}^{m'}\right)$, and $\tilde{\gamma}_{m,m'}^{(M)}$ is defined as $\tilde{\gamma}_{m,m'}^{(M)} = \left[\mathbf{P}^{-1}\right]_{m,m'} - \left[\mathbf{P}_{-n}^{-1}\right]_{m-\mathbb{1}_{\{m>n\}},m'-\mathbb{1}_{\{m'>n\}}}$.

## D INVERTIBILITY OF MATRIX $\mathbf{P}$

In this section, we first demonstrate that the empirical correlation matrix $\hat{\mathbf{P}}$ is PSD. We then show that $\mathbf{P}$ is always PD; consequently, its invertibility is guaranteed.

Let $\tilde{\boldsymbol{z}}^m \in \mathbb{R}^{|\mathcal{D}_r|d}$ denote the vectorized version of the standardized feature matrix $\tilde{\boldsymbol{Z}}^m$ for modality $m \in [M]$. According to (8), we have $[\hat{\mathbf{P}}]_{n,m} = \frac{1}{|\mathcal{D}_r|d}(\tilde{\boldsymbol{z}}^n)^T\tilde{\boldsymbol{z}}^m$. Consequently, the empirical correlation matrix $\hat{\mathbf{P}}$ can be expressed as follows:

$$
\hat{\mathbf{P}} = \frac{1}{|\mathcal{D}_r|d}\begin{bmatrix}(\tilde{\boldsymbol{z}}^1)^T\\(\tilde{\boldsymbol{z}}^2)^T\\\vdots\\(\tilde{\boldsymbol{z}}^M)^T\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{z}}^1 & \tilde{\boldsymbol{z}}^2 & \dots & \tilde{\boldsymbol{z}}^M\end{bmatrix}. \tag{15}
$$

From (15), we observe that $\hat{\mathbf{P}}$ is a Gram matrix. As shown by Horn & Johnson (2012), any Gram matrix of vectors is PSD.

Now, we show that the updated correlation matrix $\mathbf{P}$ is PD and, therefore, invertible. Let $\mathbf{P}(t)$ and $\hat{\mathbf{P}}(t)$ denote the updated correlation matrix and the empirical correlation matrix at the $t$-th update interval, respectively. Given that $\mathbf{P}$ is initialized as $\mathbf{P}^0$ and is updated via a moving average with

momentum $\beta$ at each interval $t$, we have

$$\mathbf{P}(t) = \beta^t \mathbf{P}^0 + \sum_{i=1}^{t} \beta^{t-i}(1-\beta)\hat{\mathbf{P}}(i), \quad t \geq 1. \tag{16}$$

$\mathbf{P}(t)$ in (16) is PD because, for any non-zero vector $\boldsymbol{v} \in \mathbb{R}^M$, the scalar $\boldsymbol{v}^T\mathbf{P}(t)\boldsymbol{v}$ is strictly positive. Specifically, we have

$$\begin{aligned} \boldsymbol{v}^T\mathbf{P}(t)\boldsymbol{v} &= \boldsymbol{v}^T \left[ \beta^t \mathbf{P}^0 + \sum_{i=1}^{t} \beta^{t-i}(1-\beta)\hat{\mathbf{P}}(i) \right] \boldsymbol{v} \\ &= \beta^t \left[ \boldsymbol{v}^T\mathbf{P}^0\boldsymbol{v} \right] + \sum_{i=1}^{t} \beta^{t-i}(1-\beta) \left[ \boldsymbol{v}^T\hat{\mathbf{P}}(i)\boldsymbol{v} \right]. \end{aligned} \tag{17}$$

Note that in (17), $\boldsymbol{v}^T\mathbf{P}^0\boldsymbol{v} > 0$ because $\mathbf{P}^0$ is PD, and $\boldsymbol{v}^T\hat{\mathbf{P}}(i)\boldsymbol{v} \geq 0$ because $\hat{\mathbf{P}}(i)$ is PSD. Consequently, $\boldsymbol{v}^T\mathbf{P}(t)\boldsymbol{v}$ is strictly positive, which implies that $\mathbf{P}(t)$ is PD and invertible.

## E Computational Cost Analysis of CorrCLR

Let $C_{f_n}$ and $C_{b_n}$ denote the computational costs (or FLOPs) of the forward and backward passes of the encoder corresponding to modality $n$, respectively. Similarly, let $C_{f_m}$ and $C_{b_m}$ denote the computational costs of the forward and backward passes of the encoder corresponding to modality $m$, respectively. Considering $E$ training epochs, $N$ paired samples in the training set, and a batch size of $B$, the computational cost of contrastive learning using the InfoNCE loss can be obtained as $O(EN[(C_{f_n} + C_{f_m} + C_{b_n} + C_{b_m}) + B(2d + C_{loss})])$, where $d$ is the representation dimension and $C_{loss}$ is the computational cost of computing the InfoNCE loss for each anchor sample. Since $C_{f_n}$, $C_{f_m}$, $C_{b_n}$, and $C_{b_m}$ dominate the overall computational cost, it can be approximated as $O(EN(C_{f_n} + C_{f_m} + C_{b_n} + C_{b_m}))$.

In addition to the computational cost described above, CorrCLR requires computing the correlation matrix every $T$ epochs using a subset of the training samples of size $rN$, where $r \in (0, 1]$. This process requires one forward pass for the samples from each modality. We save the representations to first compute their mean and standard deviation for feature-wise standardization, and then use the standardized representations to compute the correlation matrix. The computational cost of the forward pass and the correlation matrix calculation is $O(\frac{E}{T}rN[(C_{f_n} + C_{f_m}) + 2d^2])$, which can be approximated as $O(\frac{E}{T}rN(C_{f_n} + C_{f_m}))$.

Considering that $C_{b_n} \approx 2C_{f_n}$ and $C_{b_m} \approx 2C_{f_m}$, which is a typical approximation in deep learning (Raihan & Aamodt, 2020; Nakamura et al., 2025), CorrCLR increases the computational cost by $\frac{r}{3T} \times 100\%$. This additional computational cost can be controlled by choosing a smaller value of $r$ and a larger value of $T$. For example, when $r = 0.2$ and $T = 5$, CorrCLR increases the number of FLOPs by only $1.33\%$, which is negligible compared with the computational cost typically incurred by approaches that require hyperparameter tuning of the temperature parameter.

## F Experimental Results for the Multiview Setting

In this setting, we follow the SimCLR framework (Chen et al., 2020a), in which diverse augmented views of images are used to enhance visual representations through CL. We employ the ImageNet-1K (ILSVRC2012) dataset (Russakovsky et al., 2015), evaluating classification accuracy on the validation set after training on the full training split. Using the Lightly package (Susmelj et al., n.d.), we train a ResNet-50 from scratch. Top1 and Top5 classification accuracies are reported using $k$-NN, linear probing, and full fine-tuning as performance metrics. For baselines, we consider SimCLR (Chen et al., 2020a), MoCo v2 (Chen et al., 2020b), and InfoNCE (Oord et al., 2019) (with fixed temperatures). Additionally, we compare CorrCLR with MACL (Huang et al., 2023) and DySTreSS (Manna et al., 2025), both of which incorporate adaptive temperature strategies.

Table 7 demonstrates that CorrCLR achieves the second-best performance across all metrics by minimizing the $\rho$-InfoNCE loss in a two-view setup. Furthermore, these results show that CorrCLR can

Table 7: Comparison between CorrCLR and baseline methods in the multiview setting.

| Setup | Method | Linear Probe | | Full Fine-Tune | | $k$-NN | | Δ |
|---|---|---|---|---|---|---|---|---|
| | | Top1 | Top5 | Top1 | Top5 | Top1 | Top5 | |
| Two-View | SimCLR (Chen et al., 2020a) | 63.2 | 85.2 | 73.9 | 91.9 | 44.8 | 73.9 | 0.00 |
| | MoCo v2 (Chen et al., 2020b) | 61.5 | 84.1 | 74.3 | 91.9 | 41.8 | 72.2 | -2.07 |
| | InfoNCE (Oord et al., 2019) | 63.1 | 85.0 | 74.1 | 91.8 | 45.0 | 73.2 | -0.12 |
| | MACL (Huang et al., 2023) | 62.1 | 83.9 | 71.8 | 90.4 | 46.0 | 74.4 | -0.73 |
| | DySTreSS (Manna et al., 2025) | 65.2 | 86.9 | 74.7 | 92.4 | 49.6 | 77.9 | **+3.82** |
| | CorrCLR (Ours) | 64.6 | 86.4 | 74.6 | 92.3 | 48.7 | 76.8 | +2.94 |
| Three-View | Pairwise Alignment | 64.4 | 86.3 | 72.7 | 91.1 | 47.3 | 75.2 | +1.34 |
| | Muscle (Setayesh et al., 2026) | 66.6 | 87.7 | 75.0 | 92.8 | 51.7 | 79.3 | **+5.58** |
| | CorrCLR (Ours) | 65.8 | 87.2 | 73.9 | 91.8 | 50.2 | 78.6 | +4.13 |

be easily extended to setups with more than two views (e.g., three augmented views) by minimizing the $\rho$-Muscle loss. In contrast, existing adaptive temperature strategies, such as MACL (Huang et al., 2023) and DySTreSS (Manna et al., 2025), are designed based on the gradient behavior of the InfoNCE loss for two-view setups, making them difficult to generalize to recently proposed multimodal contrastive loss functions that effectively handle more than two views. Since the multiview setting involves a single modality, CorrCLR adaptively obtains the same correlation coefficient for each pair of views, as shown in Fig. 4.

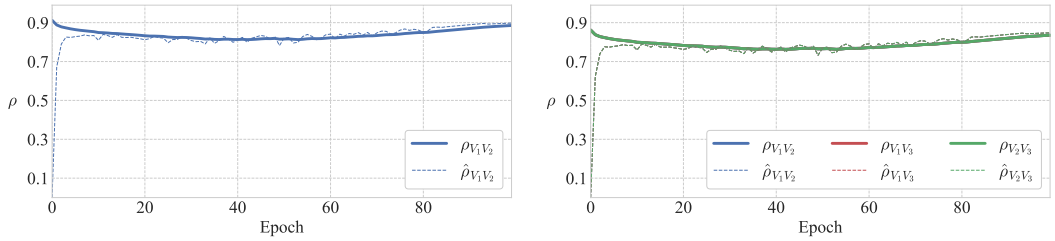



Figure 4: Evolution of updated and empirical correlation coefficients between representations of different augmented views of images in the multiview setting, for two-view (left) and three-view (right) setups. The correlation coefficients between the two views in the two-view setup and between each pair of views in the three-view setup are similar. This is consistent with the fact that representations corresponding to different augmented views belong to the same modality (i.e., images).

## G Additional Implementation Details

**3D Point-Image-Text Setting** We set the batch size $B$ and the number of training epochs $E$ to 32 and 1000, respectively. The PointBERT model is trained using the AdamW optimizer Loshchilov & Hutter (2019) with a learning rate of $5 \times 10^{-4}$ and a StepLR scheduler that decays the learning rate by 0.967 every 65600 training steps. Similarly, the SparseConv model is trained using the AdamW optimizer with a learning rate of $1 \times 10^{-3}$ and a StepLR scheduler that decays the learning rate by 0.95 every 65600 steps. All experiments under this setting were conducted on a single NVIDIA Tesla V100 GPU.

- **CorrCLR**: We set the initial correlation coefficient between each pair of modalities to 0.9.
- **ULIP** (Xue et al., 2023): We utilize the pretrained checkpoint released in the official ULIP repository.
- **OpenShape** (Liu et al., 2023): The initial temperature parameter between each pair of modalities is set to 0.07.

**Video-Audio-Text Setting** We set the batch size $B$ and the number of training epochs $E$ to 64 and 50, respectively, for all methods. We set the learning rate for EVA-CLIP-ViT-G to $5 \times 10^{-7}$ and for the rest of the models to $1 \times 10^{-4}$. We used a linear scheduler with a warmup ratio of 0.1. All experiments for this setting were conducted on eight NVIDIA Tesla V100 GPUs.

- **CorrCLR**: We set the initial correlation coefficient between each pair of modalities to 0.99.
- **Pairwise Alignment**: We set the temperature parameter $\tau_{nm} = 0.1$ in the InfoNCE loss used to align the representations of each pair of modalities $(n, m) \in \{(T, V), (T, A), (V, A)\}$.

**mmWave-WiFi-RFID Setting** We set the batch size $B$ to 64. The models are trained using the AdamW optimizer with a learning rate of $1 \times 10^{-4}$. The number of full fine-tuning epochs for the feature fusion stage is set to 40. All experiments under this setting were conducted on a single NVIDIA Tesla V100 GPU.

- **CorrCLR**: An ablation study is conducted on the initial correlation coefficients, where the values for each pair of modalities are set to 0.3, 0.6, and 0.9. The number of training epochs $E$ for CL is set to 1000.
- **Feature Fusion** (Chen & Yang, 2025): There are no CL training epochs in this method; instead, the models are fine-tuned exclusively during the feature fusion stage.
- **Muscle** (Setayesh et al., 2026): An ablation study is conducted on the temperature parameters $\tau_{nm}^{(3)}$ and $\tau_{nm}^{(2)}$ for each pair of modalities $(n, m) \in \{(M, W), (M, R), (W, R)\}$ using grid search; results are reported for the values that yield the best overall performance. The number of training epochs $E$ for CL is set to 1000.

For the baselines used across all three multimodal representation learning settings, we set the hyperparameters as follows:

- **Symile** (Saporta et al., 2024): The initial temperature parameter is set to 0.07.
- **GRAM** (Cicchetti et al., 2025b) and **TRIANGLE** (Cicchetti et al., 2025a): We set the temperature parameter to 0.07.
- **iSogCLR** (Qiu et al., 2023): We use the implemented loss function with the default values provided by the LibAUC library (Yuan et al., 2023) and set the initial temperature parameter to 0.1.

**Multiview Setting** All methods in this setting are implemented using the Lightly package (Susmelj et al., n.d.). We set the batch size $B$ and the number of training epochs $E$ to 256 and 100, respectively, for all methods. Additionally, we adopt the square root learning rate scaling recommended by SimCLR (Chen et al., 2020a) for smaller batch sizes. All experiments for this setting were conducted on two NVIDIA Tesla V100 GPUs.

- **CorrCLR**: We set the initial correlation coefficient between each pair of views to 0.99.
- **SimCLR** (Chen et al., 2020a) and **MoCo v2** Chen et al. (2020b): We use the pretrained checkpoints provided by the Lightly package (Susmelj et al., n.d.).
- **InfoNCE** (Oord et al., 2019): We set the temperature parameter $\tau$ to 0.1 in the InfoNCE loss.
- **MACL** (Huang et al., 2023): We set $\tau_0$, $\alpha$, and $\mathcal{A}_0$ to 0.1, 0.5, and 0, respectively.
- **DySTreSS** (Manna et al., 2025): We set the $\tau_{\text{min}}$ and $\tau_{\text{max}}$ to 0.1 and 0.2, respectively.
- **Pairwise Alignment**: We set the temperature parameter $\tau_{nm} = 0.1$ in the InfoNCE loss used to align the representations of each pair of views $(n, m) \in \{(1, 2), (1, 3), (2, 3)\}$.
- **Muscle** (Setayesh et al., 2026): We set the temperature parameters $\tau_{nm}^{(3)}$ and $\tau_{nm}^{(2)}$ to 0.15 and 0.1, respectively, for each pair of views $(n, m) \in \{(1, 2), (1, 3), (2, 3)\}$.

## H Correlation Coefficients Estimated by CorrCLR Using Synthetic Data

We conduct experiments to evaluate whether the multimodal correlation coefficients estimated by CorrCLR are consistent with the original correlation coefficients among multimodal data samples, as these correlations capture inherent characteristics of the data. We generate synthetic multimodal data samples using known probability distributions. The samples of each modality are passed through a 2-layer MLP to obtain modality representations, after which the $\rho$-InfoNCE loss is used to align them. We train the encoders on 50K paired samples using a batch size of 512, the Adam optimizer, and a learning rate of $1 \times 10^{-3}$ for 25 epochs. After training, we compare CorrCLR's estimated correlation $\hat{\rho}$ against the ground-truth correlation $\rho$. We also report the initial correlation coefficient

of the representations $\rho_0$, computed before CL, to isolate the effect of training from the architecture alone. We use the multivariate $t$-distribution and the multivariate Poisson distribution to generate data samples. We selected these distributions because they have closed-form expressions for the ground-truth correlation coefficients among the generated data samples.

**Multivariate $t$-Distribution** We generate a paired dataset of two synthetic modalities, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{128}$, from a multivariate $t$-distribution with location $\mathbf{0} \in \mathbb{R}^{256}$, degrees of freedom $\nu$, and scale matrix $\mathbf{\Sigma} = \mathbf{P} \otimes \boldsymbol{I}_{128}$, where $\mathbf{P}$ is a $2 \times 2$ correlation matrix with off-diagonal entry $\rho$. The ground-truth correlation coefficient between the two modalities in this setting is $\rho$. As shown in Table 8, for different values of $\rho$ and $\nu$, CorrCLR successfully estimates the correlation coefficient between the multimodal representations, and the estimated values are very close to the ground-truth correlation coefficient $\rho$.

Table 8: Comparison between the ground-truth correlation and the correlation estimated by CorrCLR for synthetically generated data samples using multivariate $t$-distribution.

| $\nu$ | $\rho$ | $\rho_0$ | $\hat{\rho}$ | $\lvert\rho - \hat{\rho}\rvert$ |
|---|---|---|---|---|
| 10 | 0.7 | 0.430 | 0.684 | 0.016 |
| 10 | 0.5 | 0.289 | 0.513 | 0.013 |
| 20 | 0.5 | 0.290 | 0.514 | 0.014 |
| 20 | 0.7 | 0.430 | 0.684 | 0.016 |

**Multivariate Poisson Distribution** We generate a paired dataset of two synthetic modalities, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{128}$, from a multivariate Poisson distribution. Specifically, each element of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ was generated independently from $\mathrm{Poisson}(\lambda_1)$ and $\mathrm{Poisson}(\lambda_2)$, respectively, with an additional shared additive component (i.e., a shared shock) for each corresponding pair of elements. The shared component was drawn from $\mathrm{Poisson}(\lambda_3)$. The ground-truth correlation coefficient between the two modalities in this setting is $\rho = \frac{\lambda_3}{\sqrt{(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)}}$. As shown in Table 9, for different values of $\lambda_1$, $\lambda_2$, and $\lambda_3$, CorrCLR estimates a correlation coefficient between the multimodal representations that is very close to the ground-truth correlation coefficient $\rho$.

Table 9: Comparison between the ground-truth correlation and the correlation estimated by CorrCLR for synthetically generated data samples using multivariate Poisson distribution.

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\rho$ | $\rho_0$ | $\hat{\rho}$ | $\lvert\rho - \hat{\rho}\rvert$ |
|---|---|---|---|---|---|---|
| 2.0 | 3.0 | 1.5 | 0.378 | 0.157 | 0.401 | 0.023 |
| 2.0 | 3.0 | 5.8 | 0.700 | 0.223 | 0.675 | 0.025 |
| 2.0 | 2.0 | 1.5 | 0.429 | 0.185 | 0.447 | 0.018 |
| 3.0 | 3.0 | 1.5 | 0.333 | 0.133 | 0.364 | 0.031 |

# I Additional Experimental Results for the Video-Audio-Text Setting

Table 10: Comparison of CorrCLR with baseline methods in the video-audio-text setting on the DiDeMo and ActivityNet datasets.

| Dataset | Method | TV | | VT | | TA | | AT | | AV | | VA | | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | |
| DiDeMo | Pairwise Alignment | 10.2 | 36.3 | 9.6 | 34.3 | 1.3 | 2.1 | 1.5 | 8.1 | 3.3 | 13.7 | 3.2 | 13.2 | 0.00 |
| | Symile (Saporta et al., 2024) | 10.8 | 38.7 | 9.9 | 35.3 | 1.6 | 8.5 | 1.5 | 8.3 | 3.5 | 15.4 | 3.3 | 14.2 | +31.5 |
| | GRAM (Cicchetti et al., 2025b) | 12.1 | 41.2 | 11.2 | 38.3 | 2.0 | 10.5 | 1.5 | 11.5 | 4.0 | 20.4 | 3.5 | 18.7 | +56.45 |
| | TRIANGLE (Cicchetti et al., 2025a) | 11.1 | 39.0 | 10.8 | 37.4 | 2.4 | 12.1 | 1.7 | 8.6 | 3.6 | 19.6 | 2.6 | 17.1 | +56.76 |
| | CorrCLR (Ours) | 12.3 | 43.6 | 11.9 | 39.3 | 2.1 | 10.8 | 1.7 | 11.7 | 4.1 | 21.0 | 3.8 | 19.2 | **+62.88** |
| ActivityNet | Pairwise Alignment | 11.3 | 41.5 | 8.3 | 37.4 | 1.9 | 6.7 | 1.2 | 5.9 | 3.2 | 12.9 | 3.4 | 14.9 | 0.00 |
| | Symile (Saporta et al., 2024) | 11.8 | 42.7 | 8.5 | 38.4 | 2.1 | 7.3 | 2.1 | 7.1 | 3.2 | 13.0 | 3.3 | 14.8 | +10.37 |
| | GRAM (Cicchetti et al., 2025b) | 15.5 | 50.1 | 12.8 | 45.7 | 1.8 | 10.7 | 1.8 | 9.1 | 3.1 | 15.9 | 3.2 | 15.7 | +26.05 |
| | TRIANGLE (Cicchetti et al., 2025a) | 19.6 | 56.9 | 18.7 | 53.1 | 2.1 | 11.3 | 1.3 | 6.3 | 2.2 | 13.5 | 1.2 | 8.5 | +19.82 |
| | CorrCLR (Ours) | 16.1 | 52.5 | 13.4 | 47.0 | 1.7 | 10.5 | 1.8 | 9.1 | 3.1 | 16.7 | 3.3 | 16.2 | **+28.22** |

# J Evolution of Correlation Coefficients During CorrCLR Training

In this section, we present figures illustrating how the correlation coefficients between multimodal representations evolve throughout the CorrCLR training procedure. Due to the interpretability of correlation coefficients, these figures enable us to identify which modality pairs are more strongly correlated. This information can further be used to understand why certain modality pairs exhibit higher correlations than others.

## J.1 Evolution of Correlation Coefficients in the Video-Audio-Text Setting

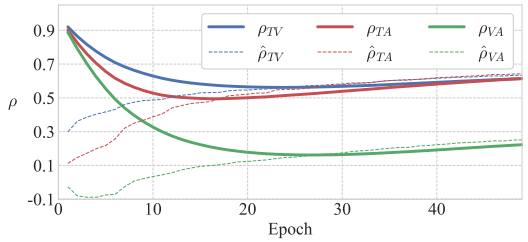



Figure 5: Evolution of updated and empirical correlation coefficients between representations of the V, A, and T modalities in the video-audio-text setting. The correlation coefficient between the V and A modalities is lower than those of the other pairs.

## J.2 Evolution of Correlation Coefficients in the mmWave-WiFi-RFID Setting

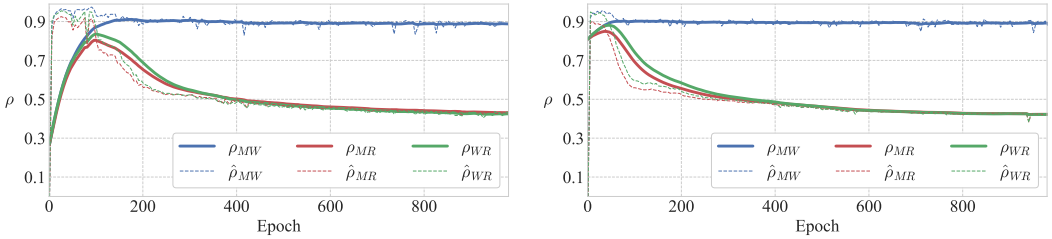



Figure 6: Evolution of updated and empirical correlation coefficients between representations of the M, W, and R modalities in the mmWave-WiFi-RFID setting for intra-domain (left) and cross-domain (right) scenarios. The correlation coefficient between the M and W modalities is much higher than those of the other pairs. Furthermore, since correlation coefficients are inherent characteristics of the data, their values do not change significantly between intra-domain and cross-domain scenarios.

# K Extended Ablation Studies

This section provides further ablation studies and extended results.

## K.1 Effect of the Initial Correlation Matrix $\mathbf{P}^0$

The results in Table 11 show that, regardless of the chosen initial correlation matrix $\mathbf{P}^0$, CorrCLR improves classification accuracy for human action recognition across various modality combinations. Furthermore, the overall performance metric $\Delta$ in Table 11 indicates that CorrCLR is relatively robust to the choice of the initial correlation matrix. In Appendix K.5, we show that the same conclusion holds when considering different initial correlation matrices $\mathbf{P}^0$ in the 3D point-image-text setting, meaning that the final correlation coefficients obtained by CorrCLR are similar across different initializations.

Table 11: Performance of CorrCLR for different initial correlation coefficients.

| $(\rho^0_{MW}, \rho^0_{MR}, \rho^0_{WR})$ | Modality | | | | | | | Δ |
|---|---|---|---|---|---|---|---|---|
| | M | W | R | M + W | M + R | W + R | M + W + R | |
| (0.3, 0.3, 0.3) | 87.0 | 84.5 | 58.1 | 91.6 | 87.1 | 83.9 | 90.3 | +28.26 |
| (0.6, 0.6, 0.6) | 87.0 | 84.8 | 55.6 | 90.1 | 86.9 | 83.8 | 90.8 | +27.43 |
| (0.9, 0.9, 0.9) | 86.7 | 84.8 | 55.5 | 88.2 | 86.4 | 84.3 | 91.0 | +27.07 |

### K.2 Effect of the Momentum Parameter $\beta$

We investigate the impact of the momentum parameter $\beta$ on the performance of CorrCLR in Fig. 7 and Table 12. A higher $\beta$ assigns more weight to the moving-averaged correlation matrix $\mathbf{P}$, while a lower $\beta$ places greater emphasis on the empirical correlation matrix $\hat{\mathbf{P}}$ during updates. As shown in Fig. 7, the evolution of $\rho_{MW}$, $\rho_{MR}$, and $\rho_{WR}$ across different $\beta$ values indicates that lower momentum values lead to faster convergence. However, because the updated correlation coefficients rely more heavily on empirical values in this case, we observe increased fluctuation. Furthermore, the final correlation coefficients in Fig. 7, along with the performance of CorrCLR in human action recognition across different $\beta$ values in Table 12, indicate that CorrCLR is largely robust to this hyperparameter.

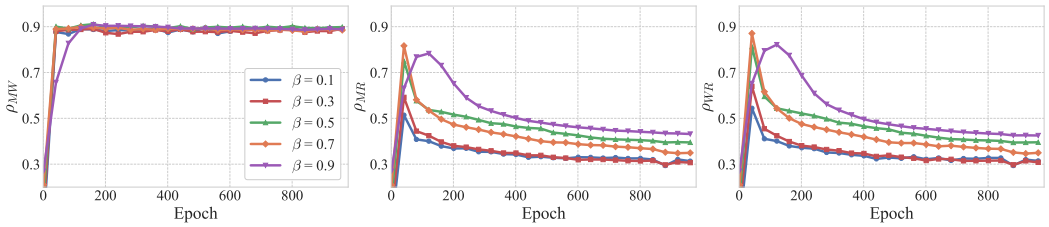



Figure 7: Evolution of $\rho_{MW}$, $\rho_{MR}$, and $\rho_{WR}$ for different momentum parameters. CorrCLR is robust to the choice of momentum parameter $\beta$.

Table 12: Effect of the momentum parameter $\beta$ on the performance of CorrCLR.

| $\beta$ | Modality | | | | | | | Δ |
|---|---|---|---|---|---|---|---|---|
| | M | W | R | M + W | M + R | W + R | M + W + R | |
| 0.1 | 87.3 | 84.5 | 57.5 | 89.6 | 86.5 | 84.4 | 90.0 | +27.72 |
| 0.3 | 87.3 | 85.0 | 56.0 | 89.9 | 85.8 | 85.0 | 90.9 | +27.61 |
| 0.5 | 87.3 | 84.2 | 55.4 | 89.1 | 87.2 | 84.7 | 90.5 | +27.35 |
| 0.7 | 87.1 | 83.9 | 54.2 | 89.9 | 85.4 | 83.8 | 89.8 | +26.19 |
| 0.9 | 87.0 | 84.5 | 58.1 | 91.6 | 87.1 | 83.9 | 90.3 | +28.26 |

### K.3 Effect of the Sampling Ratio $r$

The results in Fig. 8 show that, regardless of the chosen sampling ratio $r$, CorrCLR converges to relatively similar values of $\rho_{MW}$, $\rho_{MR}$, and $\rho_{WR}$. However, faster convergence is observed for larger values of $r$. These results, along with those in Table 4, demonstrate that CorrCLR is robust to the choice of the sampling ratio $r$.

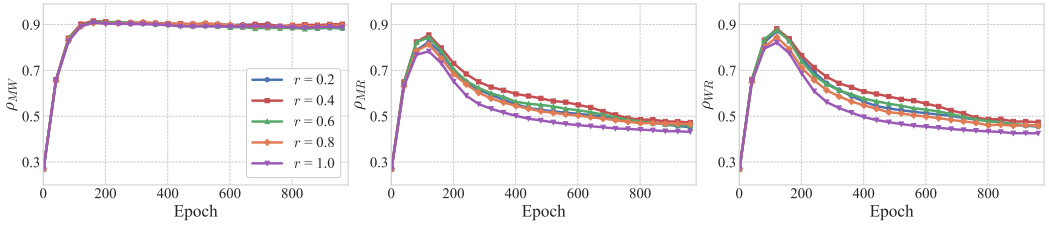



Figure 8: Evolution of $\rho_{MW}$, $\rho_{MR}$, and $\rho_{WR}$ for different sampling ratios. CorrCLR is robust to the choice of sampling ratio $r$.

### K.4 Effect of the Update Interval $T$

Table 13 shows the impact of the correlation coefficient update interval $T$ on the performance of CorrCLR. The results indicate that smaller values of $T$ (e.g., 1 and 5) yield superior performance. While a larger $T$, such as 25, still improves overall performance in terms of $\Delta$, it leads to a suboptimal solution compared to smaller intervals. Very high values, such as $T = 125$, result in decreased performance due to slow convergence. Fig. 9 illustrates the evolution of $\rho_{\text{MW}}$, $\rho_{\text{MR}}$, and $\rho_{\text{WR}}$ across different update intervals $T$. Lower values of $T$, such as $T = 1$ and $T = 5$, lead to faster convergence. Increasing $T$ to 25 not only slows convergence but also leads to a different local optimum for the correlation coefficients. While this local optimum still suggests that $\rho_{\text{MR}}$ and $\rho_{\text{WR}}$ are lower than $\rho_{\text{MW}}$, which is consistent with the results for $T = 1$ and $T = 5$, the solution obtained for $T = 25$ is suboptimal in terms of the overall performance metric $\Delta$, as shown in Table 13. Furthermore, Fig. 9 shows that within $E = 1000$ training epochs, CorrCLR fails to converge when the update interval $T$ is set to a large value such as $T = 125$. These results demonstrate that low to moderate update intervals facilitate faster convergence. Thus, we recommend choosing a small to moderate value of $T$ in CorrCLR to achieve faster convergence and higher performance. In other words, as long as $T$ is chosen such that CorrCLR converges within the specified number of training epochs, it performs well.

Table 13: Effect of the correlation coefficient update interval $T$ on CorrCLR performance.

| $T$ | Modality | | | | | | | $\Delta$ |
|---|---|---|---|---|---|---|---|---|
| | M | W | R | M + W | M + R | W + R | M + W + R | |
| 1 | 86.6 | 84.1 | 56.3 | 90.6 | 87.1 | 85.0 | 89.7 | +27.63 |
| 5 | 87.0 | 84.5 | 58.1 | 91.6 | 87.1 | 83.9 | 90.3 | +28.26 |
| 25 | 87.2 | 83.7 | 55.3 | 89.7 | 85.5 | 82.7 | 87.0 | +25.52 |
| 125 | 85.6 | 81.7 | 51.8 | 85.9 | 84.1 | 79.6 | 87.3 | +22.19 |

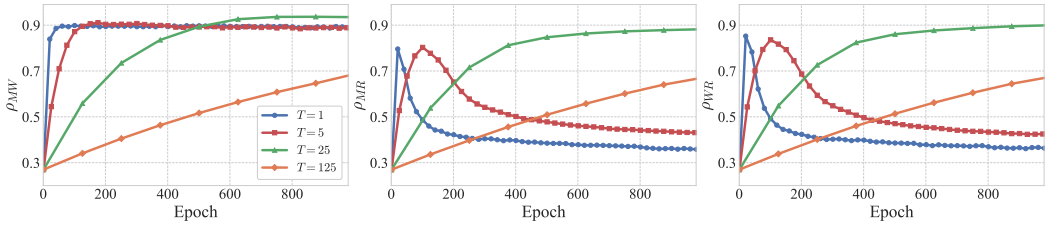



Figure 9: Evolution of $\rho_{\text{MW}}$, $\rho_{\text{MR}}$, and $\rho_{\text{WR}}$ for different correlation coefficient update intervals. Smaller update intervals, such as $T = 1$ and $T = 5$, facilitate faster convergence, whereas increasing $T$ reduces the convergence rate. Within 1000 epochs, CorrCLR has not yet converged for $T = 125$.

### K.5 Effect of $\mathbf{P}^0$ on the 3D Point-Image-Text Setting

We demonstrate the effect of different $\mathbf{P}^0$ on the final correlation coefficients obtained via CorrCLR in the 3D point-image-text setting for PointBERT and SparseConv in Figs. 10 and 11, respectively.

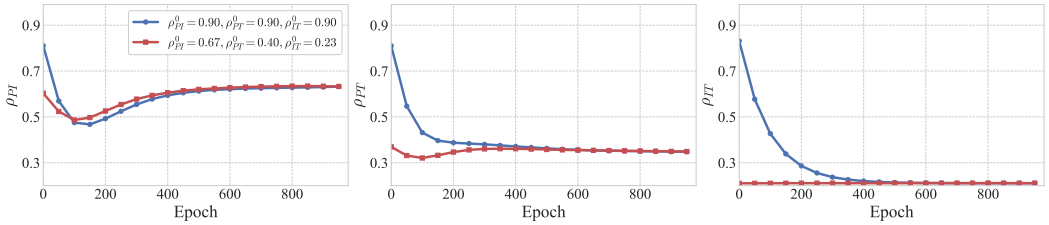



Figure 10: Evolution of $\rho_{\text{PI}}$, $\rho_{\text{PT}}$, and $\rho_{\text{IT}}$ for different initial values when PointBERT is used as the 3D encoder. CorrCLR is robust to the choice of initial correlation coefficients.

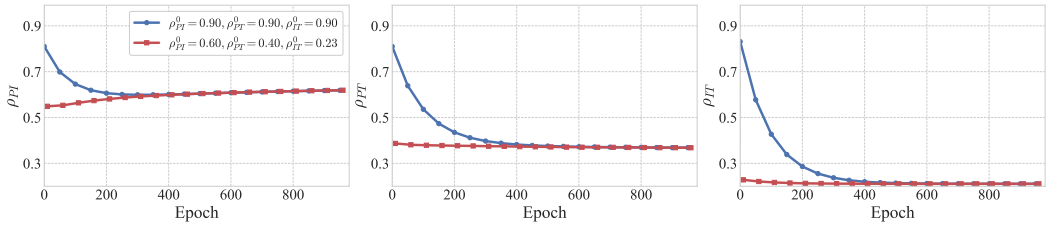



Figure 11: Evolution of $\rho_{\text{PI}}$, $\rho_{\text{PT}}$, and $\rho_{\text{IT}}$ for different initial values when SparseConv is used as the 3D encoder. CorrCLR is robust to the choice of initial correlation coefficients.

## L Multiple-Seed Tests for the Intra-Domain Scenario in the mmWave-WiFi-RFID Setting

We conduct a multi-seed experiment to evaluate the stability and robustness of CorrCLR. Specifically, the following table reports the results for the intra-domain scenario in the mmWave–WiFi–RFID setting across three different random seeds. The results in Table 14 further strengthen the reliability of our empirical claims and indicate that the improvements achieved by CorrCLR are not attributable to random variation.

Table 14: Performance comparison for the intra-domain scenario in the mmWave-WiFi-RFID setting.

| Method | Modality | | | | | | | Δ |
|---|---|---|---|---|---|---|---|---|
| | M | W | R | M + W | M + R | W + R | M + W + R | |
| Feature Fusion (Chen & Yang, 2025) | 87.8± 0.24 | 79.9± 0.15 | 54.0± 0.81 | 89.1± 2.08 | 54.7± 0.81 | 54.3± 0.63 | 55.0± 0.51 | 0.00 |
| Muscle (Setayesh et al., 2026) | 87.4± 0.57 | 84.5± 0.34 | 55.9± 0.96 | 87.2± 1.16 | 86.8± 0.40 | 83.6± 0.93 | 85.0± 0.16 | +24.84 |
| CorrCLR (Ours) | 86.3± 0.34 | 84.9± 0.13 | 57.8± 0.40 | 90.0± 0.32 | 86.8± 0.56 | 85.2± 0.98 | 89.8± 0.53 | **+27.35** |

## M Discussion on Possible Future Work Directions

**Joint Probability Distribution of Representations** In this work, we modeled the joint probability distribution of representations as a multivariate Gaussian, a choice that offers several key advantages. First, it aligns with prior studies showing that InfoNCE induces a Gaussian distribution (Betser et al., 2026), identifies the optimal lower-dimensional subspace when data samples are drawn from Gaussian mixture models (Bansal et al., 2025), and performs better when the joint probability distribution is modeled as proportional to a Gaussian kernel (Lu et al., 2023). Second, because the multivariate Gaussian distribution is closed under marginalization, the formulation remains straightforward when extending it to the general case involving more than two modalities. Third, the correlation coefficients in the multivariate Gaussian distribution facilitate the interpretation of relationships between multimodal representations and can be measured directly, as implemented in the CorrCLR algorithm.

Despite these advantages, the multivariate Gaussian distribution is not the only possible choice. Since the representations are normalized to unit length, they lie on a unit hypersphere. Consequently, several existing works have sought to establish connections between InfoNCE and distributions on the unit hypersphere (Zimmermann et al., 2021; Kirchhof et al., 2023), such as the von Mises-Fisher (vMF) distribution (Mardia & Jupp, 2009). For example, Zimmermann et al. (2021) show that InfoNCE inverts the data generation process when the marginal distribution over latent variables is uniform and the conditional distribution over positive pairs follows a vMF distribution. Additionally, the Maxwell-Poincaré-Borel lemma states that fixed-dimensional projections of the uniform distribution on high-dimensional hyperspheres converge to a standard Gaussian distribution as the dimensionality increases (Diaconis & Freedman, 1987). Exploring alternative models for the joint probability distribution of representations, their extension to cases involving more than two modalities, and their rigorous connections to the multivariate Gaussian remain promising directions for future research.

**Covariance Matrix Structure** In our formulation, we assumed that the covariance matrix of multimodal representations has the structure $\mathbf{\Sigma} = \mathbf{P} \otimes \boldsymbol{I}_d$. This implies that each modality has an isotropic covariance and that cross-modal correlations are identical across all representation dimen-

sions. This covariance structure offers several advantages. First, there is only one correlation coefficient for each pair of modalities, which enhances interpretability and clarifies how modalities are related. Second, it enables the $\rho$-InfoNCE and $\rho$-Muscle formulations to closely match the InfoNCE and Muscle loss functions, respectively. For example, since the InfoNCE loss function utilizes a single temperature parameter, $\rho$-InfoNCE is formulated using a single correlation coefficient. Moreover, the number of correlation coefficients in the $\rho$-Muscle formulation is smaller than the number of temperature parameters in the Muscle loss function. Third, the estimation of correlation coefficients in (8) is more robust because the averaging is performed over both samples and representation dimensions. However, it is possible to consider more complex structures for the covariance matrix $\mathbf{\Sigma}$. For example, one could assign distinct correlation coefficients to each representation dimension for every pair of modalities. Investigating how such extensions affect the performance of CL would be an interesting direction for future research.

**Benefits of Correlation Coefficients Learned via CorrCLR** In this work, by analyzing the evolution of correlation coefficients learned via CorrCLR across different multimodal settings, we demonstrated how these coefficients provide valuable insights into which modalities are tightly coupled and which are loosely related. In addition to the interpretability offered by these coefficients, they can provide further benefits in multimodal representation learning. For example, several recent works aim to estimate mutual information (MI) between modality representations to improve the performance of multimodal systems. MI between modalities can be used to address imbalanced cooperation in multimodal learning (Gao et al., 2025). Furthermore, it can serve as a measure of redundancy, leading to a focus on complementary modalities during training (Hadizadeh et al., 2024). Because the correlation coefficients learned by CorrCLR are related to the MI between modalities, they can be leveraged to enhance multimodal training. This represents a promising direction for future research.

Furthermore, the ability of correlation coefficients to capture negative values is another advantage of our correlation-based contrastive loss formulation when dealing with more than two modalities. In the two-modality setting, the sign of the correlation coefficient does not matter because the modality encoders (i.e., neural networks) can adjust the sign of the representations to bring positive pairs closer and push negative pairs farther apart. However, in the general case involving more than two modalities, some modalities may be positively correlated while others are negatively correlated. For example, consider forward velocity (V), optical flow magnitude (F), and altitude (H) as three modalities measured by sensors on an autonomous agricultural drone. Modalities V and F are positively correlated, whereas the correlations between H and the other two modalities are negative: at higher altitudes, the ground appears to move more slowly (lower F), and the drone often reduces speed due to control constraints (lower V). InfoNCE-based pairwise alignment or the Muscle loss function, both of which rely on purely positive temperature values, may fail to capture the true dependencies among multimodal representations in this scenario. In contrast, our correlation-based contrastive loss, which accommodates both positive and negative correlation coefficients, can successfully learn multimodal representations in such settings. Investigating the benefits of CorrCLR in scenarios involving mixed positive and negative correlations among modalities remains a promising direction for future research.

## N LIMITATIONS

- The CorrCLR algorithm can be viewed as an alternating optimization procedure, specifically an expectation-maximization (EM)-like algorithm (Dempster et al., 1977). In particular, CorrCLR first learns multimodal representations given fixed correlation coefficients. Then, given these fixed representations, a maximum likelihood estimation (MLE) problem is solved to estimate the parameters (i.e., correlation coefficients) of the multivariate Gaussian distribution. These two steps are repeated until convergence. EM-like algorithms such as CorrCLR are generally not guaranteed to converge to a global optimum (Wu, 1983). However, they can converge to a local optimum. In our extensive experiments, we showed that the resulting local optima obtained by CorrCLR consistently yield substantial improvements over state-of-the-art baselines.
- CorrCLR removes the need for heuristic temperature hyperparameter tuning in the InfoNCE and Muscle loss functions. However, it introduces four other sets of hyperparameters: the initial correlation matrix $\mathbf{P}^0$, the momentum parameter $\beta$, the sampling ratio $r$,

and the update interval $T$. At first glance, it may appear that CorrCLR merely replaces temperature tuning with another hyperparameter tuning problem; however, this is not the case. First, despite having four sets of hyperparameters, CorrCLR is robust to the choice of $\mathbf{P}^0$, $\beta$, and $r$. Furthermore, our ablation studies show that CorrCLR performs well across a range of small to moderate values for $T$. In contrast, there are no clear guidelines for selecting temperature parameters, and the performance of CL is highly sensitive to their specific values. Second, as the number of modalities increases, CorrCLR still requires only four sets of hyperparameters, whereas the number of temperature parameters in the Muscle loss grows rapidly. These factors make CorrCLR a practical algorithm for multimodal learning, regardless of the initial hyperparameter count.