



TRIAD: Ternary Information Routing for Multimodal Time-Series Sensing Signals

Yusen Liu
University of Technology Sydney
Sydney, Australia
yusen.liu@student.uts.edu.au

Shouxing Ma
University of Technology Sydney
Sydney, Australia
shouxingmaa@gmail.com

Hua Lu
Aalborg University
Copenhagen, Denmark
luhua@cs.aau.dk

Xu Cheng
Tianjin University of Technology
Tianjin, China
xu.cheng@ieee.org

Xiufeng Liu*
Technical University of Denmark
Lyngby, Denmark
xiuli@dtu.dk

Huan Huo*
University of Technology Sydney
Sydney, Australia
huan.huo@uts.edu.au

Abstract

Multimodal time series captured by heterogeneous sensors are fundamental to real-world applications, yet learning effective representations without labels remains challenging. Existing self-supervised methods adopt a binary paradigm that decomposes information into shared and private components, conflating two distinct types of shared information: redundancy, the consensus information recoverable from any single modality, and synergy, the emergent information arising only from joint observation. Most methods also rely on geometric constraints for disentanglement, which cannot guarantee semantic-level separation of information components. To address these limitations, we propose TRIAD, a framework that advances multimodal disentanglement from a binary to a ternary paradigm by separating a triad of information components: redundancy, synergy, and uniqueness. Inspired by Partial Information Decomposition theory, TRIAD employs a two-stage attention-based routing mechanism. The first stage extracts cross-modal redundancy, while the second stage disentangles synergy from modality-specific uniqueness. Predictive proxy tasks enforce the semantic definition of each component through prediction objectives. Extensive experiments on synthetic and real-world benchmarks show consistent improvements over state-of-the-art methods, with ablations, robustness tests, and interpretability analyses confirming that the learned components are functionally specialized.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Mathematics of computing** → **Time series analysis**.

Keywords

Multimodal learning, Time series, Self-supervised learning, Disentangled representation, Sensor fusion

* Corresponding authors.



This work is licensed under a Creative Commons Attribution 4.0 International License. *KDD '26, Jeju Island, Republic of Korea*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
<https://doi.org/10.1145/3770855.3818194>

ACM Reference Format:

Yusen Liu, Shouxing Ma, Hua Lu, Xu Cheng, Xiufeng Liu, and Huan Huo. 2026. TRIAD: Ternary Information Routing for Multimodal Time-Series Sensing Signals. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3770855.3818194>

Resource Availability:

The source code of this paper has been made publicly available at <https://doi.org/10.5281/zenodo.20405997>.

1 Introduction

Multimodal time series data are ubiquitous in complex real-world systems, capturing the continuous evolution of physical phenomena through heterogeneous sensing channels. Representative applications include healthcare, where ECG and respiratory signals monitor physiological states, and human activity recognition, which often relies on accelerometers and gyroscopes. The inherent data heterogeneity introduces significant modeling challenges and demands expensive expertise for annotation. Consequently, unsupervised representation learning has become essential for extracting useful features from such complex data without manual labels.

Recent advancements in self-supervised learning have focused on decomposing cross-modal information effectively. Early approaches, represented by Contrastive Learning [5, 13, 39, 40], typically operate under a unary paradigm, maximizing mutual information to extract consensus features shared across views. To capture clearer semantics, recent methods have advanced to a binary paradigm [4, 21, 22, 28], dividing the latent space into two distinct parts: modality-shared and modality-specific representations. By explicitly modeling both commonalities and individual characteristics, this binary framework has established itself as the dominant standard in current multimodal research. However, this paradigm has two fundamental limitations that hinder further progress.

First, the binary paradigm conflates synergy and redundancy by representing both within a single coarse-grained “shared” representation. Redundancy denotes consensus information that can be recovered from any individual modality, whereas synergy denotes complementary information that becomes identifiable only through joint observation and cross-modal interaction. As illustrated in Figure 1 with accelerometer and gyroscope signals for human activity recognition, both modalities support basic discrimination between

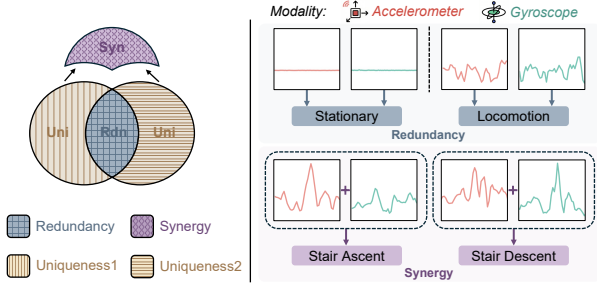


Figure 1: Left: ternary disentanglement in TRIAD. Right: an HAR example illustrating redundancy (stationary vs locomotion) and synergy (stair ascent vs descent) using two modalities (accelerometer and gyroscope sensing signals).

static and dynamic states. However, more fine-grained distinctions, such as level walking versus walking upstairs, require complementary cues, where the accelerometer reflects linear impact forces and the gyroscope captures body inclination. When these qualitatively different factors are collapsed into one shared latent variable, the resulting representation becomes ambiguous and difficult to interpret. Moreover, objectives based on similarity maximization are inherently biased toward capturing redundancy, that is, the intersection of information across modalities. Consequently, existing models do not explicitly encode synergy, which limits the ability to model cross-modal interactions that are essential for a holistic characterization of the underlying system.

Second, the methodology for disentanglement relies heavily on geometric constraints that are fundamentally limited. To achieve disentanglement, prevalent approaches force latent vectors to be orthogonal by minimizing their dot product or cosine similarity. However, geometric orthogonality primarily regulates pairwise linear correlation in the learned coordinate system and does not provide a functional definition of what each component is expected to encode. As a result, representations can appear well separated geometrically while remaining entangled at the semantic level. This mismatch makes it difficult to obtain components that are semantically interpretable and consistent with information-theoretic notions such as redundancy, synergy, and uniqueness.

In this work, we bridge these gaps by drawing inspiration from Partial Information Decomposition theory [43], which provides a principled basis for decomposing multimodal information into redundancy, synergy, and uniqueness. We observe that the semantic definitions of these three components can be operationalized through proxy tasks that directly encode their information-theoretic roles. Based on this insight, we propose TRIAD, an information routing framework that advances multimodal disentanglement from a binary to a ternary paradigm. Leveraging the naturally predictive structure of sensing signals, TRIAD adopts task-defined disentanglement, where the meaning of each component is enforced through prediction objectives inspired by PID theory. This ternary paradigm is realized through a two-stage attention-based routing mechanism. In Stage 1, cross-modal attention identifies information that remains predictively stable across arbitrary modality subsets and routes it to the redundancy component. In Stage 2, the non-redundant residual is further decomposed: information that yields improved predictions when all modalities are jointly observed is

routed to synergy, while information that supports prediction only within its own modality is routed to uniqueness. This hierarchical design mitigates information leakage between components through sequential decomposition. Our main contributions are as follows:

- **Ternary Paradigm:** We propose TRIAD, a framework that advances multimodal learning from a binary to a ternary paradigm. To the best of our knowledge, it is the first to separate synergy from redundancy and uniqueness for multimodal sensing signals.
- **Hierarchical Routing:** We design a two-stage attention-based gated routing mechanism that extracts redundancy before separating synergy from uniqueness, minimizing information leakage via sequential decomposition.
- **Task-Defined Disentanglement:** We introduce predictive proxy tasks that enforce functional separation by operationalizing the semantic definition of each component, offering a robust alternative to geometric constraints.
- **Comprehensive Evaluation:** Extensive experiments on synthetic and real-world benchmarks demonstrate consistent improvements over baselines. Ablations, robustness tests, and interpretability analyses further validate component specialization.

2 Related Work

2.1 Self-Supervised Learning for Time Series

Self-supervised learning (SSL) for time series extracts representations from unlabeled data by exploiting intrinsic structural information [37, 48]. These approaches have largely evolved along two lines: contrastive learning and reconstruction-based learning [16, 49]. Early instance-discrimination frameworks such as SimCLR [6, 17, 31] inspired a range of time-series adaptations [10, 14, 30, 45, 49, 51]. TNC [40] exploits temporal neighborhoods by treating nearby windows as positives to enforce temporal continuity. TS-TCC [13] combines temporal and contextual contrast to improve robustness to augmentations and context shifts. TS2Vec [47] learns universal representations via hierarchical contrastive learning across multiple temporal resolutions. CoST [44] further introduces contrastive learning for disentangling seasonal-trend components. Beyond the time domain, TF-C [50], FAT [7], FrERA [38], and BTSF [46] incorporate frequency-domain cues to strengthen contrastive representations. AutoTCL [52] and InfoTS [29] design automatic augmentations to better preserve temporal semantics. TimeDRL [5] improves representation via disentangled dual-level embeddings.

In addition to contrastive learning, reconstruction-based objectives have been explored. SimMTM [9] proposes a masked time-series modeling framework for pre-training. PITS [23] adopts independent patch embeddings with a lightweight reconstruction objective. Most time-series SSL methods target single-modality inputs and primarily capture temporal structure shared across augmentations or masking patterns. In real-world sensing systems, heterogeneous sensors exhibit complementarity and cross-modal interactions, driving interest in multimodal time-series SSL.

2.2 Multimodal Representation Learning for Sensing Signals

Multimodal representation learning integrates information from multiple modalities into a representation for downstream tasks

while accounting for cross-modal alignment and modality-specific cues [12, 15, 20, 25–27, 34, 41, 42]. In the visio-language domain, large-scale contrastive pretraining has proven effective: CLIP [34] demonstrates that aligning image-text pairs induces strong zero-shot transfer, ALIGN [19] further scales such contrastive alignment, and BLIP-2 [24] bridges modality gaps with improved efficiency.

Recently, multimodal learning has attracted increasing attention in sensing applications, where heterogeneous sensors are treated as modalities [8, 32]. Representative approaches adapt contrastive multiview learning to maximize cross-modal agreement, e.g., Contrastive Multiview Coding (CMC) [39], and further improve consistency via explicit geometric constraints in a shared latent space [33]. Many recent methods follow a binary decomposition into shared and modality-specific factors. FOCAL [28] factorizes sensing representations into shared and private components with orthogonality-based regularization. FreqMAE [21] introduces a frequency-aware masked autoencoder for scalable sensing pretraining. MATE [4] proposes disentangled learning for multimodal sensing signals. InfoMAE [22] focuses on pair-efficient cross-modal alignment.

However, existing methods for multimodal time-series sensing often adopt a binary paradigm (shared vs. private) that conflates redundancy with synergy. FactorCL [25] already separates information into more than two parts, but it keeps redundancy and synergy in one shared component. Similarity-based objectives bias learning toward redundancy (information recoverable from individual modalities), while leaving synergy (information available only through joint observation) under-modeled. To the best of our knowledge, TRIAD is among the first multimodal time-series frameworks to isolate synergistic information as a distinct latent component, advancing the dominant binary paradigm to a ternary decomposition aligned with Partial Information Decomposition.

3 Preliminaries: Information-Theoretic Foundations

This section presents information-theoretic foundations that motivate the proposed ternary decomposition. The section formalizes the learning problem, introduces Partial Information Decomposition as a conceptual reference, and derives task-based characterizations aligned with the predictive objectives used in TRIAD.

3.1 Problem Formulation

Consider a multimodal time series $\mathbf{X} \in \mathbb{R}^{T \times C}$ comprising M modalities, where T denotes the sequence length and $C = \sum_{m=1}^M C_m$ denotes the total number of channels. The input is partitioned into a history window $\mathbf{X}_h \in \mathbb{R}^{T_h \times C}$ and a prediction horizon $\mathbf{X}_f \in \mathbb{R}^{T_p \times C}$, where $T = T_h + T_p$. An encoder f_θ maps the history window to three sets of latent representations:

$$f_\theta : \mathbf{X}_h \mapsto (\mathbf{z}_R, \mathbf{z}_S, \{\mathbf{z}_U^{(m)}\}_{m=1}^M), \quad (1)$$

where $\mathbf{z}_R \in \mathbb{R}^{N \times d_R}$ encodes redundant information, $\mathbf{z}_S \in \mathbb{R}^{N \times d_S}$ encodes synergistic information, and $\mathbf{z}_U^{(m)} \in \mathbb{R}^{N \times d_U}$ encodes information that is unique to modality m . Here N denotes the number of temporal tokens after patchification.

3.2 Partial Information Decomposition

Partial Information Decomposition (PID) [3, 43] provides a conceptual reference for separating redundant, unique, and synergistic information. Given two sources X_1, X_2 and a target Y , PID decomposes the mutual information into four non-negative components:

$$I(X_1, X_2; Y) = R + U_1 + U_2 + S. \quad (2)$$

These components are defined by accessibility patterns: redundancy (R) is accessible from each source individually, uniqueness (U_1 or U_2) is accessible only from a single source, and synergy (S) is accessible only under joint observation. In this work, the modality-specific history observations $X_h^{(1)}, \dots, X_h^{(M)}$ serve as sources and the future observation $Y = X_f$ serves as the target. We adopt future prediction of the target as our proxy task, which provides a direct and naturally available self-supervision signal for sensing time series.

3.3 Predictive Operationalization of PID

This work operationalizes the accessibility patterns in Eq. (2) through predictive comparisons that match the training objectives. Let $\Omega \subseteq \{1, \dots, M\}$ denote a set of visible modalities, where Z^Ω denotes a representation inferred from $\{X_h^{(m)}\}_{m \in \Omega}$. Let $\mathcal{Q}$ denote a decoder family of conditional predictive distributions, where $\mathcal{Q} = \{q_\theta(y | z) : \theta \in \Theta\}$. We define the optimal conditional log-loss risk as

$$C_Q(Y | Z) \triangleq \inf_{q \in \mathcal{Q}} \mathbb{E}[-\log q(Y | Z)], \quad (3)$$

where the infimum is taken over all $q \in \mathcal{Q}$. Based on this risk, we define the task-induced information as $I_Q(Z \rightarrow Y) \triangleq -C_Q(Y | Z)$. The theoretical justification of this operationalization results from the inequality

$$I(Z; Y) \geq H(Y) - C_Q(Y | Z), \quad (4)$$

which holds when $H(Y)$ is finite. Eq. (4) implies that improving prediction within $\mathcal{Q}$ increases a certified lower bound on $I(Z; Y)$ without assuming loss-MI equivalence [2]. We therefore treat the task-induced information as a predictive quantity rather than an estimate of the mutual information itself, and TRIAD optimizes predictive risks instead of estimating PID terms.

The semantic PID patterns translate into risk relations. Redundancy is operationalized by requiring that predictive risk is stable across non-empty visibility sets up to a margin:

$$C_Q(Y | Z_R^\Omega) \leq C_Q(Y | Z_R^{\Omega'}) + \epsilon, \quad \forall \Omega, \Omega' \subseteq \{1, \dots, M\}, |\Omega|, |\Omega'| \geq 1, \quad (5)$$

where Z_R^Ω denotes the redundant representation inferred from visibility set Ω and $\epsilon \geq 0$ is a margin. Synergy is operationalized by a strict joint-observation gain:

$$C_Q(Y | Z_S^{\text{all}}) < C_Q(Y | Z_S^\Omega), \quad \forall \Omega \subsetneq \{1, \dots, M\}, \quad (6)$$

where Z_S^{all} denotes the synergistic representation inferred from all modalities. Uniqueness is operationalized as asymmetric decodability under an adversarial decoder family, where $C_Q(Y_m | Z_U^{(m)})$ is small and $C_{Q_{\text{adv}}}(Y_{-m} | Z_U^{(m)})$ is large, Y_m and Y_{-m} denote the target components of modality m and of all other modalities, respectively, and Q_{adv} denotes an adversarial decoder family. Section 4.5 instantiates Eqs. (5)–(6) as training objectives through margin-based comparisons and adversarial suppression.

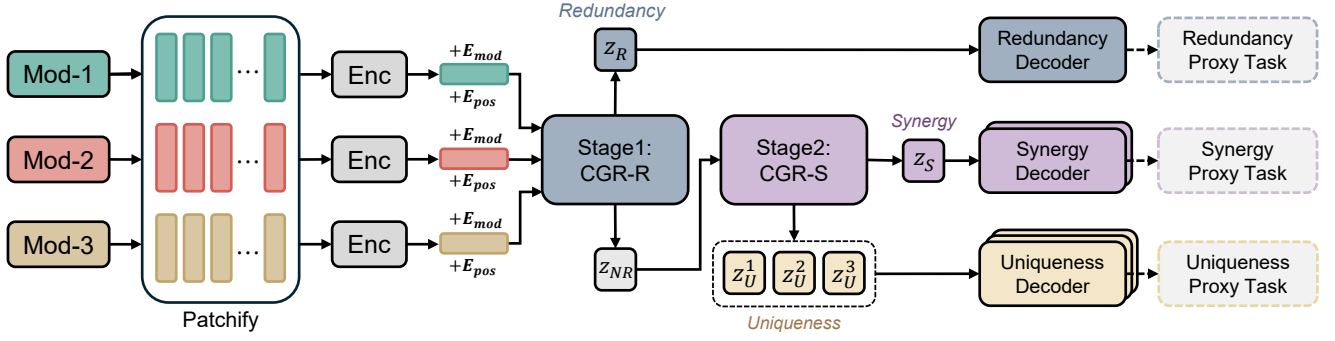


Figure 2: Overall architecture of TRIAD. The framework hierarchically decomposes multimodal time series into redundancy (z_R), synergy (z_S), and uniqueness ($z_U^{(m)}$) via two-stage cross-modal gated routing (CGR) and proxy tasks. The pipeline consists of (1) patchified temporal encoding with per-modality backbones, (2) two-stage CGR routing that extracts redundancy and then separates synergy from uniqueness, and (3) task-defined decoders and predictive proxy tasks for each component.

4 Methodology

We present TRIAD, which realizes the ternary decomposition via two innovations: a two-stage attention-based routing mechanism for hierarchical disentanglement, and predictive proxy tasks that operationalize the characterizations detailed in Section 3.3.

4.1 Overall Architecture

Figure 2 shows the architecture of TRIAD, which has three modules: (1) **Patchified Temporal Encoding** maps each modality to token sequences; (2) **Two-Stage CGR Routing** decomposes information into redundancy, synergy, and uniqueness via cross-modal gating (Figure 3); and (3) **Task-Defined Decoders** enforce disentanglement via predictive proxy tasks (Figure 4).

Given the history window $X_h \in \mathbb{R}^{T_h \times C}$, we partition it into M modality-specific views $\{X^{(m)}\}_{m=1}^M$ and apply patchification and modality-specific temporal encoding. The resulting tokens pass through two CGR stages: Stage 1 extracts redundancy shared across modalities, and Stage 2 decomposes the residual into synergy and uniqueness. Finally, dedicated decoders predict future values from each component, and proxy tasks enforce their intended semantics.

4.2 Patchified Temporal Encoding

To model local temporal structure efficiently, we tokenize each modality via patchification to reduce sequence length. For each modality m , the history window $X^{(m)} \in \mathbb{R}^{T_h \times C_m}$ is partitioned into $N = T_h/s$ non-overlapping patches (assuming s divides T_h):

$$P^{(m)} = \text{Patchify}(X^{(m)}) \in \mathbb{R}^{N \times (C_m \cdot s)}. \quad (7)$$

Each patch is treated as a token and encoded by a modality-specific temporal backbone of 1D convolutions with residual connections:

$$H^{(m)} = \text{TemporalBackbone}_m(P^{(m)}) \in \mathbb{R}^{N \times D}, \quad (8)$$

where D denotes the hidden dimension. To preserve temporal order and modality identity, we add learnable position embeddings $E_{\text{pos}} \in \mathbb{R}^{N \times D}$ and a modality embedding $E_{\text{mod}} \in \mathbb{R}^D$ (broadcast across the N tokens), followed by layer normalization:

$$Z^{(m)} = \text{LayerNorm}(H^{(m)} + E_{\text{pos}} + E_{\text{mod}}). \quad (9)$$

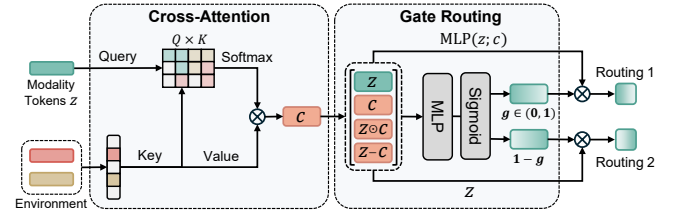


Figure 3: Cross-Modal Gated Routing (CGR). Each modality queries a leave-one-out environment through cross-attention, and a learned gate routes token-level information into routed and residual branches.

4.3 Cross-Modal Gated Routing

The core module of TRIAD is Cross-Modal Gated Routing (CGR). As shown in Figure 3, CGR uses cross-modal attention to retrieve contextual evidence and a token-wise gate to allocate information across latent branches. The module is instantiated twice with stage-specific parameters: CGR-R extracts redundancy, and CGR-S further decomposes the residual into synergy and uniqueness.

4.3.1 Leave-One-Out Environment Construction. For each modality m , we construct a leave-one-out environment $\mathcal{E}^{(-m)}$ by concatenating token sequences from all modalities except m . This provides cross-modal context while keeping query and context tokens explicitly separated:

$$E^{(-m)} = \text{Concat}(\{Z^{(j)}\}_{j \neq m}) \in \mathbb{R}^{(M-1)N \times D}, \quad (10)$$

$$\hat{E}^{(-m)} = \text{MLP}(E^{(-m)}). \quad (11)$$

We apply a lightweight MLP to $E^{(-m)}$ to obtain a stronger context sequence; the resulting tokens $\hat{E}^{(-m)}$ serve as memory for subsequent cross-modal attention.

4.3.2 Cross-Modal Attention and Gated Routing. For each modality m , token representations $Z^{(m)}$ query the leave-one-out environment via multi-head cross-attention to retrieve the most relevant cross-modal evidence:

$$C^{(m)} = \text{LayerNorm}(\text{Attn}(Z^{(m)}, \hat{E}^{(-m)}, \hat{E}^{(-m)})), \quad (12)$$

where $\text{Attn}(\cdot)$ denotes scaled dot-product attention with $\mathbf{Z}^{(m)}$ as queries and $\hat{\mathbf{E}}^{(-m)}$ as keys and values. The context sequence $\mathbf{C}^{(m)}$ summarizes token-level evidence from other modalities predictive for modality m . We then compute a soft gate from interaction features between $\mathbf{Z}^{(m)}$ and $\mathbf{C}^{(m)}$:

$$\mathbf{g}^{(m)} = \sigma \left(\text{MLP} \left([\mathbf{Z}^{(m)}; \mathbf{C}^{(m)}; \mathbf{Z}^{(m)} \odot \mathbf{C}^{(m)}; \mathbf{Z}^{(m)} - \mathbf{C}^{(m)}] \right) \right), \quad (13)$$

where σ is the sigmoid function, $\odot$ denotes element-wise multiplication, and $[\cdot; \cdot]$ denotes concatenation. This design captures raw signals, agreement, and residual discrepancies, providing a compact summary of cross-modal compatibility. Finally, the gate routes information into two branches:

$$\mathbf{F}^{(m)} = \text{MLP} \left([\mathbf{Z}^{(m)}; \mathbf{C}^{(m)}] \right), \quad (14)$$

$$\mathbf{Z}_{\text{route-1}}^{(m)} = \mathbf{g}^{(m)} \odot \mathbf{F}^{(m)}, \quad (15)$$

$$\mathbf{Z}_{\text{route-2}}^{(m)} = (1 - \mathbf{g}^{(m)}) \odot \mathbf{Z}^{(m)}. \quad (16)$$

The gate $\mathbf{g}^{(m)} \in (0, 1)^{N \times 1}$ implements token-wise soft routing. Since CGR is used in multiple stages, *route-1* and *route-2* are stage-agnostic; their semantics are determined by the stage definition.

4.4 Two-Stage Hierarchical Decomposition

The key insight of TRIAD is that multimodal information benefits from hierarchical decomposition. Stage 1 isolates redundancy, i.e., information shared across modalities and recoverable from any non-empty modality subset. Stage 2 operates on the residual to separate synergy from uniqueness. This sequential formulation reduces ambiguity and mitigates leakage by identifying synergy and uniqueness only after removing redundancy.

4.4.1 Stage 1: Redundancy Extraction. In Stage 1, the routing module with parameters θ_R decomposes each modality token sequence into a redundancy-bearing part and a non-redundant residual. Given tokens $\mathbf{Z}^{(m)}$ and the leave-one-out environment $\mathcal{E}^{(-m)}$, it outputs:

$$\mathbf{Z}_R^{(m)}, \mathbf{Z}_{NR}^{(m)} = \text{CGR}_R(\mathbf{Z}^{(m)}, \mathcal{E}^{(-m)}; \theta_R), \quad (17)$$

where $\mathbf{Z}_R^{(m)}$ are tokens routed to the redundancy branch and $\mathbf{Z}_{NR}^{(m)}$ are retained as the non-redundant residual for Stage 2. The final redundancy representation projects $\mathbf{Z}_R^{(m)}$ to dimension d_R and averages across modalities:

$$\mathbf{z}_R = \frac{1}{M} \sum_{m=1}^M \mathbf{W}_R \mathbf{Z}_R^{(m)}, \quad \mathbf{W}_R \in \mathbb{R}^{D \times d_R}, \quad (18)$$

where $\mathbf{W}_R$ is a learnable projection matrix.

4.4.2 Stage 2: Synergy and Uniqueness Separation. The residual tokens $\{\mathbf{Z}_{NR}^{(m)}\}$ contain information not captured as redundancy, including (i) synergy predictive only under joint observation and (ii) modality-specific information. To separate these factors, we apply a second routing module with independent parameters θ_S :

$$\mathbf{Z}_S^{(m)}, \mathbf{Z}_U^{(m)} = \text{CGR}_S(\mathbf{Z}_{NR}^{(m)}, \mathcal{E}_{-m}^{NR}; \theta_S). \quad (19)$$

where $\mathcal{E}_{-m}^{NR}$ is the leave-one-out environment from the non-redundant tokens of all modalities except m . In this stage, information inferred from other modalities via cross-attention is routed to $\mathbf{Z}_S^{(m)}$ (synergy), while remaining modality-specific content is routed to $\mathbf{Z}_U^{(m)}$ (uniqueness).

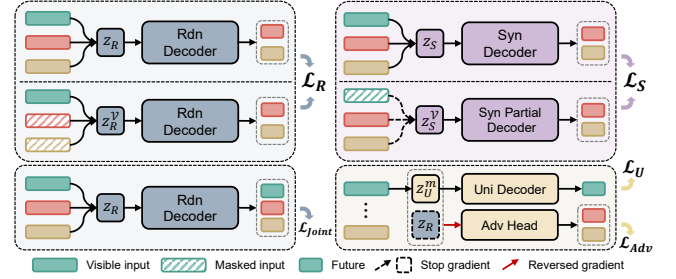


Figure 4: Task-defined disentanglement via predictive proxy tasks. Each proxy task enforces functional separation among redundancy, synergy, and uniqueness components.

Synergy is aggregated across modalities, while uniqueness remains modality-specific:

$$\mathbf{z}_S = \frac{1}{M} \sum_{m=1}^M \mathbf{W}_S \mathbf{Z}_S^{(m)}, \quad \mathbf{z}_U^{(m)} = \mathbf{W}_U^{(m)} \mathbf{Z}_U^{(m)}, \quad (20)$$

where $\mathbf{W}_S \in \mathbb{R}^{D \times d_S}$ and $\mathbf{W}_U^{(m)} \in \mathbb{R}^{D \times d_U}$ are learnable projection matrices. Mean pooling here only summarizes the synergy tokens.

4.5 Proxy Tasks and Training Objective

As illustrated in Figure 4, the ternary decomposition in TRIAD is enforced through predictive proxy tasks that instantiate the risk relations in Section 3.3. The training objectives are defined in terms of conditional predictive risk, which provides a task-defined operationalization of predictively accessible information.

For each sample, we sample a visible subset $\mathcal{V} \subset \{1, \dots, M\}$ with $\emptyset \neq \mathcal{V} \neq \{1, \dots, M\}$ and denote its complement by $\mathcal{V}^c$. Modalities in $\mathcal{V}^c$ are zeroed out during encoding, yielding partial representations $\mathbf{z}_R^{\mathcal{V}}$ and $\mathbf{z}_S^{\mathcal{V}}$. We use the following notation. Let $\mathbf{Y} = \text{Patchify}(\mathbf{X}_f)$ be the patchified future target and $\hat{\mathbf{Y}}$ be predictions. For these $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, subscripts index modalities ($\mathbf{Y}_{\mathcal{V}}, \mathbf{Y}_{\mathcal{V}^c}$) and superscripts index observation conditions of input ($\hat{\mathbf{Y}}^{(\text{all})}, \hat{\mathbf{Y}}^{(\mathcal{V})}$).

4.5.1 Joint Prediction. The joint objective encourages the redundant representation to remain broadly predictive of the future across modalities. A shared decoder predicts the patchified future:

$$\hat{\mathbf{Y}}^{(\text{all})} = \text{Decoder}_R(\mathbf{z}_R), \quad \mathcal{L}_{\text{Joint}} = \|\hat{\mathbf{Y}}^{(\text{all})} - \mathbf{Y}\|_2^2. \quad (21)$$

4.5.2 Redundancy: Cross-Modal Recovery under Partial Observation. Redundancy is operationalized by a subset-invariance constraint on predictive risk (Eq. (5)). For a sampled visibility set $\mathcal{V}$, we evaluate the redundancy decoder on the masked channels $\mathcal{V}^c$ under partial observation:

$$\hat{\mathbf{Y}}^{(\mathcal{V})} = \text{Decoder}_R(\mathbf{z}_R^{\mathcal{V}}). \quad (22)$$

The redundancy objective enforces that partial observation does not degrade masked-channel prediction beyond a margin:

$$\mathcal{L}_R = \max \left(0, \|\hat{\mathbf{Y}}_{\mathcal{V}^c}^{(\mathcal{V})} - \mathbf{Y}_{\mathcal{V}^c}\|_2^2 - \|\hat{\mathbf{Y}}_{\mathcal{V}^c}^{(\text{all})} - \mathbf{Y}_{\mathcal{V}^c}\|_2^2 - \epsilon_{\text{red}} \right), \quad (23)$$

where $\epsilon_{\text{red}} \geq 0$ is a margin. This objective encourages $\mathbf{z}_R^{\mathcal{V}}$ to preserve predictively accessible information stable across visibility sets.

4.5.3 Synergy: Predictive Gain from Joint Observation. Synergy is operationalized by a strict joint-observation gain (Eq. (6)). The comparison is performed on the same visible channels $\mathcal{V}$ to avoid a trivial advantage from observing additional inputs. Full-observation prediction uses $\mathbf{z}_S$, while partial-observation prediction uses a Synergy Partial Decoder and does not update the encoder through the partial branch:

$$\hat{\mathbf{Y}}^{(\text{all})} = \text{Decoder}_S(\mathbf{z}_S), \quad \hat{\mathbf{Y}}^{(\mathcal{V})} = \text{Decoder}_S^{\mathcal{V}}(\text{sg}(\mathbf{z}_S^{\mathcal{V}})). \quad (24)$$

The prediction errors on visible channels are

$$\ell_S^{\text{all}} = \|\hat{\mathbf{Y}}_{\mathcal{V}}^{(\text{all})} - \mathbf{Y}_{\mathcal{V}}\|_2^2, \quad \ell_S^{\mathcal{V}} = \|\hat{\mathbf{Y}}_{\mathcal{V}}^{(\mathcal{V})} - \mathbf{Y}_{\mathcal{V}}\|_2^2. \quad (25)$$

The synergy objective combines predictive terms with a margin constraint that requires a strict improvement under full observation:

$$\mathcal{L}_S = \ell_S^{\text{all}} + \ell_S^{\mathcal{V}} + \max(0, \ell_S^{\text{all}} - \text{sg}(\ell_S^{\mathcal{V}}) + \epsilon_{\text{syn}}), \quad (26)$$

where $\epsilon_{\text{syn}} \geq 0$ is a margin. This objective encourages $\mathbf{z}_S$ to encode predictively accessible information that yields a strict joint-observation gain relative to partial baselines.

4.5.4 Uniqueness: Self-Prediction with Conditional Adversarial Regularization. Uniqueness is operationalized as asymmetric decodability under a conditional adversary. Each modality-specific decoder predicts its modality future from its unique representation:

$$\hat{\mathbf{Y}}^{(m)} = \text{Decoder}_{U,m}(\mathbf{z}_U^{(m)}), \quad \mathcal{L}_U = \frac{1}{M} \sum_{m=1}^M \|\hat{\mathbf{Y}}^{(m)} - \mathbf{Y}^{(m)}\|_2^2. \quad (27)$$

To suppress cross-modal leakage, a conditional adversarial decoder predicts the future of the remaining modalities from the same unique representation, and the encoder increases the adversarial error. The adversarial decoder is conditioned on $\mathbf{z}_R$ with stop-gradient to test leakage beyond redundancy:

$$\hat{\mathbf{Y}}^{(-m)} = \text{Head}_{\text{adv},m}([\text{sg}(\mathbf{z}_R), \text{GRL}(\mathbf{z}_U^{(m)})]), \quad (28)$$

$$\mathcal{L}_{\text{Adv}} = \frac{1}{M} \sum_{m=1}^M \|\hat{\mathbf{Y}}^{(-m)} - \mathbf{Y}^{(-m)}\|_2^2, \quad (29)$$

where $\mathbf{Y}^{(-m)}$ denotes all modalities except m and $\text{GRL}(\cdot)$ denotes the gradient reversal layer. The adversarial head minimizes $\mathcal{L}_{\text{Adv}}$, while gradient reversal causes the encoder to maximize $\mathcal{L}_{\text{Adv}}$, thus suppressing the best achievable cross-modal predictability from $\mathbf{z}_U^{(m)}$ under the adversarial head family.

4.5.5 Overall Objective. The full pretraining objective combines joint prediction with the three disentanglement objectives:

$$\mathcal{L} = \mathcal{L}_{\text{Joint}} + \gamma_R \mathcal{L}_R + \gamma_S \mathcal{L}_S + \gamma_U \mathcal{L}_U + \eta \mathcal{L}_{\text{Adv}}, \quad (30)$$

where γ_R , γ_S , γ_U , and η are scalar weights.

5 Experiments

We conduct extensive experiments to assess TRIAD across a diverse suite of multimodal time series benchmarks. The experimental design aims to answer the following research questions.

- **RQ1 (Disentanglement Validity):** Can TRIAD recover ground-truth latent factors in controlled synthetic settings?
- **RQ2 (Representation Quality):** Does TRIAD outperform state-of-the-art self-supervised methods under multiple evaluation protocols on real-world datasets?

Table 1: Dataset statistics.

Dataset	Domain	#Classes	#Mod	Seq. Len	#Samples
SynTime	Synthetic	4	2	200	15000
HAR	HAR	6	9	128	10299
PAMAP2	HAR	12	4	256	2724
D1NAMO	Health	3	5	300	4479
SDH	Building	2	4	256	25926
WESAD	Affect	4	6	256	18148

Table 2: SynTime results on Class Redundancy (Class 1), Class Synergy (Class 4) and Overall. Mean over 5 seeds.

Method	Class Rdn		Class Syn		Overall	
	Acc	F1	Acc	F1	Acc	F1
SimCLR	99.17	99.17	95.06	95.19	98.48	98.49
CMC	100.00	100.00	86.66	92.72	96.67	98.18
GMC	100.00	100.00	86.06	92.42	96.52	98.11
TNC	99.89	99.92	92.67	94.79	98.14	98.68
TS-TCC	97.94	96.92	89.11	90.53	96.46	96.44
TimeDRL	56.56	57.79	83.39	71.31	84.28	81.61
FOCAL	100.00	100.00	62.61	60.62	90.65	90.16
FreqMAE	98.39	99.19	74.44	63.83	93.21	90.56
InfoMAE	63.94	65.26	35.00	33.56	74.65	71.51
MATE	100.00	100.00	66.67	44.43	91.35	79.17
TRIAD	100.00	100.00	99.61	99.75	99.90	99.94

- **RQ3 (Component Analysis):** Do the disentangled components (R/S/U) exhibit functional specialization and remain consistent with their intended semantic definitions?
- **RQ4 (Robustness):** Do the learned representations remain stable under missing modalities and variations in key hyperparameters?
- **RQ5 (Qualitative Analysis):** Can qualitative interpretability analyses provide evidence that each component captures the intended information type?

5.1 Experimental Setup

5.1.1 Datasets. We evaluate TRIAD on one synthetic dataset, **SynTime**, and five real-world multimodal time series datasets from diverse domains (Table 1). SynTime exposes ground-truth redundancy, uniqueness, and synergy factors; its construction is described in Appendix A.1. Real-world benchmarks include **HAR** [1] and **PAMAP2** [35] for human activity recognition, **D1NAMO** [11] and **WESAD** [36] for physiological sensing, and **SDH** [18] for building sensing. Details are provided in Appendix A.2.

5.1.2 Baselines. We compare TRIAD with twelve baselines across four categories. Contrastive Learning: **SimCLR** [6]. Multimodal Self-Supervised Learning: **CMC** [39], **GMC** [33], **FactorCL** [25], **InfMasking** [42]. Temporal Self-Supervised Learning: **TNC** [40], **TS-TCC** [13], **TimeDRL** [5]. Multimodal Temporal Self-Supervised Learning: **FOCAL** [28], **FreqMAE** [21], **MATE** [4], **InfoMAE** [22]. All baselines were tuned under the same validation protocol, using official code when available. Following prior state-of-the-art studies, we report macro-F1 and accuracy for all datasets.

5.2 Synthetic Results (RQ1)

To isolate multimodal information patterns, we evaluate on SynTime, a two-modality synthetic benchmark with four classes: redundancy (Class 1), modality-specific cues (Class 2 and Class 3),

Table 3: Linear probing results (Accuracy % and F1-score). Best in bold, second-best underlined. Mean over 5 seeds.

Method	HAR		PAMAP2		D1NAMO		SDH		WESAD		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SimCLR	65.73	66.21	64.35	58.98	59.31	43.62	87.36	65.64	50.10	38.04	65.37	54.50
CMC	87.17	87.57	73.91	74.40	60.34	43.13	92.98	83.80	78.26	63.88	78.53	70.56
GMC	87.48	87.98	83.77	84.52	62.33	53.26	<u>93.00</u>	84.10	77.95	65.13	80.91	75.00
TNC	65.32	65.55	69.28	70.47	58.99	54.71	<u>86.43</u>	60.89	64.99	43.35	69.00	58.99
TS-TCC	81.44	82.03	78.84	78.83	59.77	52.17	91.90	81.56	66.97	52.07	75.78	69.33
TimeDRL	82.39	82.79	88.12	88.86	61.29	60.74	88.24	72.62	70.08	59.93	78.02	72.99
FOCAL	83.07	83.41	72.75	71.76	63.08	58.76	89.67	78.36	<u>80.38</u>	75.24	77.79	73.51
FreqMAE	85.82	85.48	<u>91.01</u>	<u>91.42</u>	<u>72.03</u>	65.27	90.36	75.90	72.57	58.67	82.36	75.35
InfoMAE	86.66	87.26	<u>89.27</u>	90.07	69.09	64.51	91.55	81.48	74.22	60.91	82.16	76.85
MATE	90.50	90.16	89.86	90.69	68.82	67.31	90.75	81.34	74.84	65.28	82.95	78.96
FactorCL	<u>91.96</u>	<u>91.80</u>	90.72	90.86	49.66	33.67	87.10	77.08	78.70	74.39	79.63	73.56
InfMasking	90.46	90.28	79.71	82.23	60.25	25.90	90.45	79.11	68.09	52.20	77.79	65.94
TRIAD	92.95	92.87	93.62	91.86	75.46	70.55	93.57	85.88	84.80	81.23	88.08	84.48

Table 4: k -NN ($k=5$) probing results. Mean over 5 seeds.

Method	HAR		SDH		WESAD	
	Acc	F1	Acc	F1	Acc	F1
CMC	85.48	85.93	90.83	78.24	72.42	60.91
GMC	84.70	84.93	91.33	<u>79.20</u>	<u>72.60</u>	60.94
TS-TCC	68.04	68.55	89.93	78.83	62.76	44.44
TimeDRL	82.08	82.86	87.74	66.75	64.67	44.97
FreqMAE	75.03	74.49	87.74	66.75	71.28	<u>65.16</u>
InfoMAE	78.83	78.71	84.67	45.85	66.36	48.93
MATE	84.42	84.79	87.00	69.01	69.78	54.44
TRIAD	89.75	89.43	<u>90.88</u>	79.92	76.13	70.82

and synergy (Class 4), identifiable only via joint observation. Table 2 reports results for Class 1, Class 4, and overall. On Class 1, most methods are near the ceiling, whereas Class 4 varies widely, from 35.00%/33.56% (InfoMAE) to 99.61%/99.75% (TRIAD) in Accuracy/F1. Notably, classic methods score higher on Class 4 (SimCLR: 95.06%/95.19%; TS-TCC: 89.11%/90.53%) than recent binary-decomposition baselines, suggesting the latter under-capture or suppress synergy. TRIAD achieves the best Class 4 result and overall performance. Full per-class results are in Appendix B.1.

5.3 Real-World Results (RQ2)

5.3.1 Evaluation Protocols. To assess representation quality, we employ three evaluation protocols. **Linear Probing:** We freeze the pretrained encoder and train a linear classifier on top of frozen representations. **k -NN Evaluation:** We perform k -nearest neighbors voting ($k = 5$) on frozen representations to assess neighborhood structure without parameter updates. **Fine-tuning:** We fine-tune the model end-to-end with a classification head and unfreeze the last two encoder layers.

5.3.2 Linear Probing Results. Table 3 shows that TRIAD achieves the best linear-probing performance on all five datasets. On average, TRIAD attains 88.08% accuracy and 84.48% F1-score, outperforming the second-best method by 5.13 and 5.52 points, corresponding to relative improvements of 6.2% and 7.0%, respectively. Gains are especially pronounced on datasets with heterogeneous physiological modalities: WESAD (+4.42 points over FOCAL) and D1NAMO (+3.43 points over FreqMAE), where cross-modal interactions encode complementary dynamics. On motion-centric benchmarks, TRIAD also improves HAR (+0.99 over FactorCL) and PAMAP2

Table 5: Fine-tuning results. Mean over 5 seeds.

Method	HAR		SDH		WESAD	
	Acc	F1	Acc	F1	Acc	F1
CMC	90.60	91.04	92.86	84.82	86.67	84.96
GMC	90.87	91.27	92.31	82.38	84.14	83.12
TS-TCC	87.28	87.16	94.14	87.53	83.14	80.69
TimeDRL	88.19	88.66	92.95	83.79	82.26	77.06
FreqMAE	<u>94.37</u>	<u>94.23</u>	<u>94.81</u>	89.05	84.39	80.56
InfoMAE	91.75	91.41	93.90	86.93	80.38	76.36
MATE	92.47	92.69	91.43	82.99	83.80	81.04
TRIAD	96.71	96.75	95.33	89.89	87.88	85.58

Table 6: Ablations study. Gray indicates Acc (%) drop.

Variant	HAR	SDH	WESAD
TRIAD	92.95	93.57	84.80
w/o Residual	92.05 (-0.90)	88.50 (-5.07)	77.52 (-7.28)
w/o Hierarchy	90.97 (-1.98)	89.66 (-3.91)	78.41 (-6.39)
w/o Attention	91.94 (-1.01)	89.19 (-4.38)	80.57 (-4.23)
w/o Gating	90.55 (-2.40)	88.75 (-4.82)	76.04 (-8.76)
w/o $\mathcal{L}_{\text{Joint}}$	90.13 (-2.82)	88.15 (-5.42)	69.45 (-15.35)
w/o $\mathcal{L}_U$	86.52 (-6.43)	89.20 (-4.37)	70.06 (-14.74)
w/o $\mathcal{L}_R$	90.57 (-2.38)	89.21 (-4.36)	75.06 (-9.74)
w/o $\mathcal{L}_S$	87.82 (-5.13)	89.40 (-4.17)	68.97 (-15.83)
w/o $\mathcal{L}_{\text{Adv}}$	90.97 (-1.98)	89.02 (-4.55)	77.23 (-7.57)

(+2.61 over FreqMAE). After Bonferroni correction, TRIAD remains significant on all datasets ($p < 0.05$), with a 2.40-point average accuracy gain over the second-best baseline.

5.3.3 k -NN and Fine-Tuning Results. In Table 4, TRIAD performs best on HAR (89.75%/89.43%) and WESAD (76.13%/70.82%), with gains of +4.27 and +3.53 accuracy points over the second-best. On SDH, TRIAD reaches 90.88% accuracy, slightly below GMC (91.33%), while attaining the best macro F1-score (79.92% vs. 79.20% for GMC), indicating better class-balanced neighborhood quality.

In Table 5, TRIAD is best across all three datasets: 96.71%/96.75% on HAR, 95.33%/89.89% on SDH, and 87.88%/85.58% on WESAD. Margins over the strongest baselines remain positive after end-to-end adaptation, including a 2.34 point accuracy gain on HAR over FreqMAE (96.71% vs. 94.37%) and a 1.21 point accuracy gain on WESAD over CMC (87.88% vs. 86.67%). Together with the k -NN results, these trends indicate the pretrained representations support both local neighborhood consistency and downstream fine-tuning.

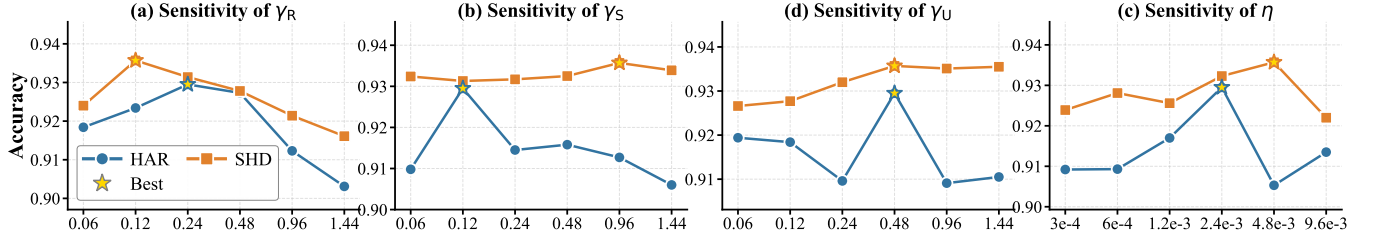


Figure 5: Hyperparameter sensitivity analysis.

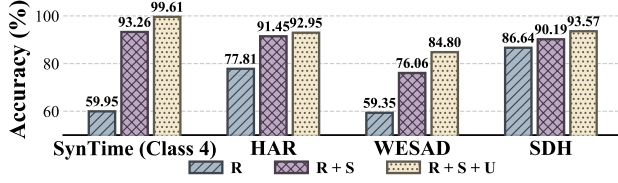


Figure 6: Representation composition analysis.

5.4 Component Analysis (RQ3)

5.4.1 Ablation Study. To validate our hierarchical design, we conduct ablations under linear probing. Table 6 summarizes the degradation when specific modules or objective functions are removed.

The **w/o Residual** variant feeds Stage 2 the original features instead of the nonredundant residual, reducing accuracy by 7.28 points on WESAD and 5.07 points on SDH. The **w/o Hierarchy** variant performs a one-shot ternary split, degrading accuracy on WESAD (6.39 points) and SDH (3.91 points). These results indicate that passing a residual into Stage 2 and staged decomposition both improve separation between redundancy and the remaining factors. Replacing token-wise cross-attention with a global context (**w/o Attention**) reduces WESAD accuracy by 4.23 points and SDH accuracy by 4.38 points, suggesting token-level alignment matters beyond a global summary. Removing adaptive gating (**w/o Gating**) yields the largest drop, reducing WESAD accuracy by 8.76 points and SDH accuracy by 4.82 points, indicating sample-adaptive routing is important when the proportions of redundant and interaction-driven information change over time.

Objective ablations in Table 6 show WESAD is the most sensitive dataset. Removing $\mathcal{L}_S$ yields the largest drop (15.83 points), and removing $\mathcal{L}_{\text{Joint}}$ causes a comparable degradation (15.35 points). Removing $\mathcal{L}_U$ also reduces performance, especially on WESAD (14.74 points) and HAR (6.43 points). Removing $\mathcal{L}_R$ and $\mathcal{L}_{\text{Adv}}$ further drops WESAD accuracy by 9.74 and 7.57 points. Together, these results suggest each term helps limit information leakage and maintain a non-trivial allocation of redundant, synergistic, and unique factors. A capacity-controlled ablation is provided in Appendix B.2.

5.4.2 Hyperparameter Sensitivity Analysis. We study sensitivity to four key hyperparameters, γ_R , γ_S , γ_U , and η . Figure 5 summarizes results on HAR and SDH. For γ_R , performance peaks at 0.24 on HAR (92.95%) and 0.12 on SDH (93.57%); overly small values under-enforce redundancy, while overly large values over-constrain the representation and degrade accuracy. For γ_S , HAR performs best at 0.12 and SDH at 0.96, suggesting the synergy constraint is more dataset dependent, although SDH remains relatively stable across

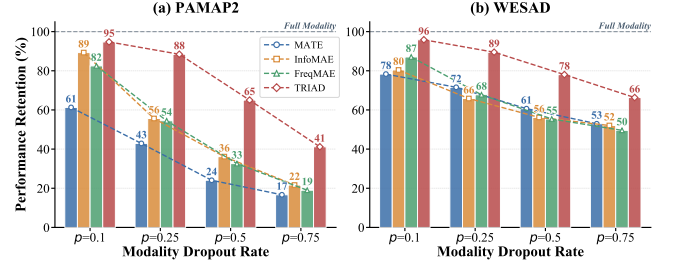


Figure 7: Robustness to modality dropout. TRIAD degrades gradually compared to the sharp drops observed in baselines.

the range. For γ_U , both datasets peak at 0.48 (HAR: 92.95%, SDH: 93.57%), and SDH remains stable from 0.24 to 1.44. For η , best performance is achieved at 2.4×10^{-3} on HAR and 4.8×10^{-3} on SDH, indicating a typical trade-off between convergence speed and optimization stability. Overall, performance is stable across a reasonable range, with accuracy varying within 2–3% in most cases. Additional margin sensitivity analyses are provided in Appendix B.4.

5.4.3 Representation Composition Analysis. To assess functional specialization, we train linear probes on subsets of the learned representations while fixing the classifier architecture and input dimensionality; unused components are set to zero so performance differences reflect information content rather than dimensionality. Figure 6 shows consistent gains from adding the synergy component. On SynTime (Class 4), synergy improves performance by 33.31% over redundancy-only, indicating that the synergy branch captures complementary information beyond redundancy. On HAR, WESAD, and SDH, adding synergy improves accuracy by 13.64, 16.71, and 3.55 points, respectively. Overall, using redundancy, synergy, and uniqueness together performs best.

5.5 Robustness Analysis (RQ4)

To evaluate robustness to sensor failure, we simulate missing modalities at test time via random modality dropout with rates $p \in \{0.1, 0.25, 0.5, 0.75\}$. Figure 7 reports the retention rate, i.e., downstream performance normalized to the full-modality setting, and shows that TRIAD maintains higher retention than the baselines as modality dropout increases. On PAMAP2, TRIAD retains about 65% when half the modalities are dropped ($p = 0.5$), whereas InfoMAE and FreqMAE retain only about 36% and 33%, respectively. Under the most severe dropout ($p = 0.75$), TRIAD remains above 40%, while the baselines drop below 25%. On WESAD, the retention advantage is larger: TRIAD preserves about 78% of full-modality

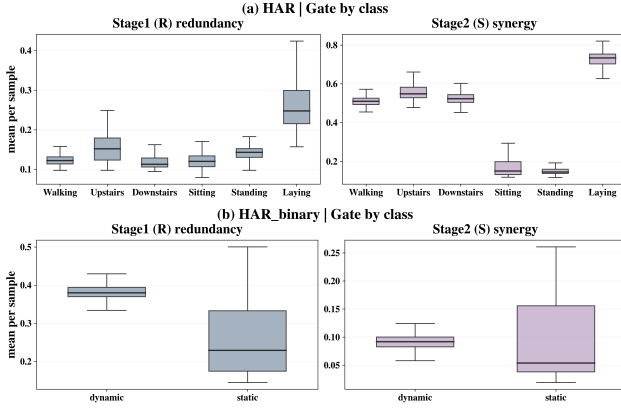


Figure 8: Class-wise distributions of gating coefficients.

performance at $p = 0.5$, while the baselines retain roughly 55–61%. At $p = 0.75$, TRIAD remains above 65%, compared with about 50–53% for the baselines. These results suggest that TRIAD’s ternary disentanglement yields stronger robustness to missing modalities than the binary disentanglement baselines.

5.6 Case Study: Task-Dependent Routing (RQ5)

To assess whether the routed factors align with downstream requirements, we run a case study on HAR. A coarse-grained variant, HAR_binary, is formed by merging the six activity classes into two super-classes: Dynamic (Walking, Upstairs, Downstairs) and Static (Sitting, Standing, Lying). Inputs are identical; only label granularity changes. This isolates how the utility of redundancy (z_R) and synergy (z_S) varies with task granularity.

5.6.1 Class-Wise Routing Behavior. Figure 8 summarizes class-wise routing behavior by presenting the distributions of the gating coefficients. On HAR (a), Stage 2 gates are higher for dynamic activities, with a substantial fraction of samples exceeding 0.5, which indicates that fine-grained discrimination among these dynamic classes relies on synergy information. On HAR_binary (b), Stage 1 gates are consistently higher than Stage 2 gates, and Stage 2 gates remain concentrated near zero for both classes. This pattern indicates that the coarse-grained dynamic–static separation is supported primarily by redundant information shared across sensors.

5.6.2 Gating Visualization. Figure 9 plots mean gating coefficients across test samples for Stage 1 and Stage 2. Larger values indicate stronger routing to the corresponding branch. Figure 9 (a) and (b) show the redundancy gate shifting toward larger values on HAR_binary, indicating increased routing to redundancy when the task is dominated by shared motion cues. Figure 9 (c) and (d) show the complementary pattern: HAR activates the synergy gate more, whereas HAR_binary concentrates near zero. This suggests that coarse-grained labels primarily rely on cross-modal redundancy, while finer-grained recognition benefits from cross-modal synergy.

6 Limitations

While TRIAD establishes a new ternary paradigm for multimodal sensing, we acknowledge three limitations defining its scope. First, the PID-inspired components are learned through predictive proxy

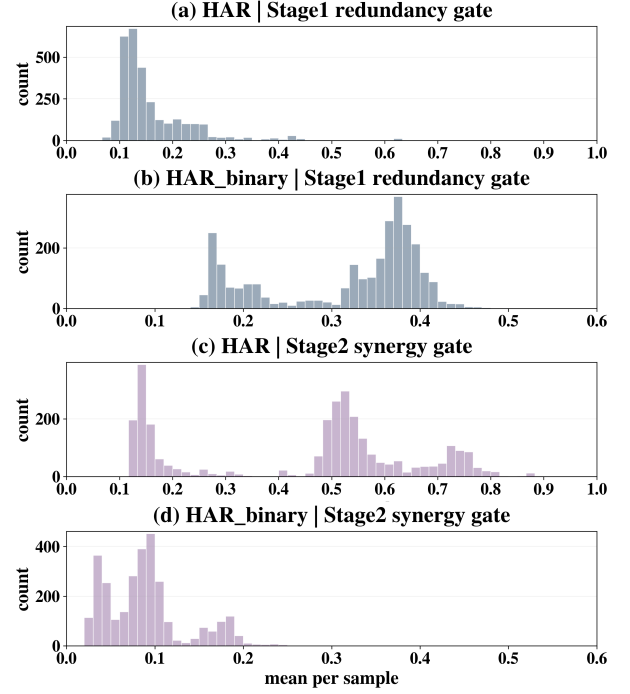


Figure 9: Learned two-stage gating coefficient distributions, showing marked shifts between HAR and HAR_binary.

tasks rather than exact information-theoretic estimation, and thus should be interpreted as task-defined predictive components. Second, the current attention-based routing assumes temporally synchronized modalities and may require adaptation for asynchronous sensing scenarios. Finally, fine-grained disentanglement introduces additional computational overhead, making the model slightly less efficient than simpler unary baselines.

7 Conclusion

We present TRIAD, a self-supervised framework that advances multimodal time series learning from binary shared/private decomposition to a ternary paradigm separating redundancy, synergy, and uniqueness. Inspired by PID theory and the inherent predictability of time series, TRIAD employs two-stage attention-based routing with task-defined predictive objectives to enforce the functional role of each component. Experiments on synthetic and five real-world benchmarks demonstrate consistent improvements over recent baselines; ablation studies validate the effectiveness of each component, while case studies and interpretability analyses confirm that TRIAD achieves disentanglement aligned with PID semantics. Future work will explore efficient routing architectures and training objectives supporting partial-modality scenarios.

Acknowledgments

This work was supported by the Marie Skłodowska-Curie Postdoctoral Individual Fellowship (Grant No. 101154277), the RE-INTEGRATE project funded by the European Union’s Horizon Europe Research and Innovation Programme (Grant No. 101118217), and the China Scholarship Council (Grant No. 202208410132).

References

- [1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. 2013. A public domain dataset for human activity recognition using smartphones. In *Esann*, Vol. 3. 3–4.
- [2] David Barber and Felix Agakov. 2004. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems* 16, 320 (2004), 201–208.
- [3] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. 2014. Quantifying unique information. *Entropy* 16, 4 (2014), 2161–2183.
- [4] Ruichu Cai, Zhifan Jiang, Kaitao Zheng, Zijian Li, Weilin Chen, Xuexin Chen, Yifan Shen, Guangyi Chen, Zhifeng Hao, and Kun Zhang. 2025. Learning disentangled representation for multi-modal time-series sensing signals. In *Proceedings of the ACM on Web Conference 2025*. 3247–3266.
- [5] Ching Chang, Chiao-Tung Chan, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. 2024. Timedr: Disentangled representation learning for multivariate time-series. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 625–638.
- [6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [7] Rui Cheng, Xiangfei Jia, Qing Li, Rong Xing, Jiwen Huang, Yu Zheng, and Zhilong Xie. 2025. Fat: Frequency-aware pretraining for enhanced time-series representation learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 310–321.
- [8] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. 2022. Cocoa: Cross modality contrastive learning for sensor data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–28.
- [9] Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Ming-sheng Long. 2023. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems* 36 (2023).
- [10] Jufang Duan, Wei Zheng, Yangzhou Du, Wenfa Wu, Haipeng Jiang, and Hongsheng Qi. 2024. MF-CLR: Multi-frequency contrastive learning representation for time series. In *Forty-first International Conference on Machine Learning*.
- [11] Fabien Dubosson, Jean-Eudes Ranvier, Stefano Bromuri, Jean-Paul Calbimonte, Juan Ruiz, and Michael Schumacher. 2018. The open DiNAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. *Informatics in Medicine Unlocked* 13 (2018), 92–100.
- [12] Benoît Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. 2025. What to align in multimodal contrastive learning?. In *International Conference on Learning Representations*, Vol. 2025. 5408–5432.
- [13] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. Time-series representation learning via temporal and contextual contrasting. *International Joint Conference on Artificial Intelligence* (2021).
- [14] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. 2019. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems* 32 (2019).
- [15] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. 2022. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204* (2022).
- [16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.
- [17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.
- [18] Dezhi Hong, Jorge Ortiz, Kamin Whitehouse, and David Culler. 2013. Towards automatic spatial verification of sensor placement in buildings. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*. 1–8.
- [19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.
- [20] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2023. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7661–7671.
- [21] Denizhan Kara, Tomoyoshi Kimura, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, and Tarek Abdelzaher. 2024. Freqmae: Frequency-aware masked autoencoder for multi-modal iot sensing. In *Proceedings of the ACM Web Conference 2024*. 2795–2806.
- [22] Tomoyoshi Kimura, Xinlin Li, Osama Hanna, Yatong Chen, Yizhuo Chen, Denizhan Kara, Tianshi Wang, Jinyang Li, Xiaomin Ouyang, Shengzhong Liu, et al. 2025. Infomae: Pair-efficient cross-modal alignment for multimodal time-series sensing signals. In *Proceedings of the ACM on Web Conference 2025*. 3084–3095.
- [23] Seunghan Lee, Taeyoung Park, and Kibok Lee. 2024. Learning to Embed Time Series Patches Independently. In *The Twelfth International Conference on Learning Representations*.
- [24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [25] Paul Pu Liang, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. In *Advances in Neural Information Processing Systems*.
- [26] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. 2023. High-Modality Multimodal Transformer: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning. *Transactions on Machine Learning Research* (2023).
- [27] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. 2023. M3AE: Multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 1657–1665.
- [28] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. 2023. Focal: Contrastive learning for multivariate time-series sensing signals in factorized orthogonal latent space. *Advances in Neural Information Processing Systems* 36 (2023), 47309–47338.
- [29] Dongsheng Luo, Wei Cheng, Yingheng Wang, Dongkuan Xu, Jingchao Ni, Wen-chao Yu, Xuchao Zhang, Yanchi Liu, Yuncong Chen, Haifeng Chen, and Xiang Zhang. 2023. Time series contrastive learning with information-aware augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4534–4542.
- [30] Qianwen Meng, Hangwei Qian, Yong Liu, Lizhen Cui, Yonghui Xu, and Zhiqi Shen. 2023. MHCL: masked hierarchical cluster-wise contrastive learning for multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9153–9161.
- [31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
- [32] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 324–337.
- [33] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. 2022. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*. PMLR, 17782–17800.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [35] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmark dataset for activity monitoring. In *2012 16th international symposium on wearable computers*. IEEE, 108–109.
- [36] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 400–408.
- [37] Zineb Senane, Lele Cao, Valentin Leonhard Buchner, Yusuke Tashiro, Lei You, Paweł Andrzej Herman, Mats Nordahl, Ruibo Tu, and Vilhelm Von Ehrenheim. 2024. Self-supervised learning of time series representation via diffusion process and imputation-interpolation-forecasting mask. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2560–2571.
- [38] Tian Tian, Chunyan Miao, and Hangwei Qian. 2025. FreRA: a frequency-refined augmentation for contrastive learning on time series classification. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 2835–2846.
- [39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *European conference on computer vision*. Springer, 776–794.
- [40] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. 2021. Unsupervised representation learning for time series with temporal neighborhood coding. *International Conference on Learning Representations* (2021).
- [41] Xinming Tu, Zhi-Jie Cao, Sara Mostafavi, Ge Gao, et al. 2022. Cross-linked unified embedding for cross-modality representation learning. *Advances in Neural Information Processing Systems* 35 (2022), 15942–15955.
- [42] Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang, Jun Wang, Zenglin Xu, and Jiang Duan. 2026. InfMasking: Unleashing Synergistic Information by Contrastive Multimodal Interactions. *Advances in Neural Information Processing Systems* 38 (2026), 15529–15555.
- [43] Paul L Williams and Randall D Beer. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515* (2010).
- [44] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. 2022. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *International Conference on Learning Representations* (2022).

- [45] Yuhan Wu, Xiyu Meng, Yang He, Junru Zhang, Haowen Zhang, Yabo Dong, and Dongming Lu. 2024. Multi-view Self-Supervised Contrastive Learning for Multivariate Time Series. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 9582–9590.
- [46] Ling Yang and Shenda Hong. 2022. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In *International conference on machine learning*. PMLR, 25038–25054.
- [47] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8980–8987.
- [48] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2114–2124.
- [49] Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, et al. 2024. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE transactions on pattern analysis and machine intelligence* 46, 10 (2024), 6775–6794.
- [50] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems* 35 (2022), 3988–4003.
- [51] Xiaochen Zheng, Xingyu Chen, Manuel Schürch, Amina Mollaysa, Ahmed Allam, and Michael Krauthammer. 2023. Simts: Rethinking contrastive representation learning for time series forecasting. *arXiv preprint arXiv:2303.18205* (2023).
- [52] Xu Zheng, Tianchun Wang, Wei Cheng, Aitian Ma, Haifeng Chen, Mo Sha, and Dongsheng Luo. 2024. Parametric Augmentation for Time Series Contrastive Learning. *International Conference on Learning Representations* (2024).

A Dataset Details

A.1 Synthetic Dataset Construction

SynTime is a dual-modality synthetic time series classification benchmark with controlled ground-truth factors for redundancy, uniqueness, and synergy. Each sample is a length-200 sequence with two channels, one per modality. The dataset is class-balanced across training, validation, and test splits with 10,000, 2,000, and 3,000 samples, respectively. All samples undergo per-sample, per-modality z-score normalization over time to suppress energy- and variance-based shortcuts, forcing models to rely on temporal structure and cross-modal relationships. The dataset contains four classes, each defined by a distinct generative process (Class 1–Class 4):

Class 1 (Redundancy). A shared latent process $u(t)$ is an Ornstein-Uhlenbeck (OU) process with sinusoidal perturbations. The two modalities receive different nonlinear projections: modality A observes $\tanh(u + b_1) + \epsilon_1$, while modality B observes $u^3 + b_2 + \epsilon_2$. Thus, the consensus information is not linearly observable and requires nonlinear alignment.

Class 2 and Class 3 (Uniqueness). These classes contain modality-specific information. In Class 2, only modality A contains a structured signal (a nonlinearly transformed OU-sinusoidal process), while modality B is noise; Class 3 reverses these roles.

Class 4 (Synergy). The two modalities are generated by coupled Lorenz systems with coupling strength $K = 5.0$. Although each modality is chaotic, the discriminative information lies in the joint dynamics and synchronization patterns, requiring both modalities to be observed together.

For integration, the Lorenz systems use a time step of $dt = 0.01$ and retain every fifth sample. A burn-in period ensures that trajectories reach the attractor before sampling. This design provides a controlled testbed for evaluating whether a representation learning method can disentangle redundancy, uniqueness, and synergy. In particular, Class 4 serves as a ground-truth synergy class, where

accurate classification requires cross-modal interactions that are inaccessible from either modality alone.

A.2 Real-World Dataset Details

The benchmarks span diverse sensing tasks, modalities, and label spaces. We report accuracy (the fraction of correctly predicted samples) and macro-F1 (the unweighted mean of per-class F1 scores).

HAR [1]. This dataset contains smartphone inertial recordings from 30 subjects performing six activities, sampled at 50 Hz. We treat each axis of total acceleration, body acceleration, and angular velocity as one modality, resulting in nine modalities.

PAMAP2 [35]. This dataset includes 18 activities from 9 subjects with wearable heart-rate and IMU signals (temperature, accelerometer, gyroscope, magnetometer, and orientation). We use the Protocol recordings, retain 12 classes, and treat heart rate and the three IMUs (hand, chest, ankle) as four modalities.

D1NAMO [11]. This wearable diabetes-monitoring dataset contains ECG, respiration, accelerometer, and glucose measurements from 29 participants. We construct per-second mean and standard deviation features from ECG, respiration, and 3-axis accelerometer signals, define three glucose-based classes (hypoglycemia/in-range/hyperglycemia), and treat ECG, respiration, and each accelerometer axis as five modalities.

SDH [18]. This building-sensing dataset provides environmental measurements, including CO₂, temperature, humidity, luminosity, and PIR motion sensor data, collected from 51 rooms in an office building. The preprocessing uses CO₂, humidity, temperature, and light as inputs and derives binary occupancy labels from PIR. In our modality grouping, each input channel (CO₂, humidity, temperature, light) is regarded as one modality, resulting in four modalities.

WESAD [36]. This dataset focuses on stress and affect classification. The preprocessing uses chest-worn signals (3-axis ACC, ECG, EDA, EMG, respiration, and temperature) to distinguish four affective states. Under the modality grouping used in this work, ACC is treated as one modality, and each of the other five signals is treated as one modality; overall, the grouping results in six modalities.

B Additional Analysis

B.1 Detailed Synthetic Results

We report the full per-class results for the synthetic dataset SynTime in Table 7. Performance on Class 1–Class 3 (redundancy and uniqueness) is near saturation for most methods, whereas performance on Class 4 (synergy) varies substantially.

Binary-decomposition baselines show a larger performance drop on Class 4 than on Class 1–Class 3. This pattern is consistent across variants: FOCAL performs well on Class 1–Class 3 but drops to 62.61/60.62 on Class 4, while masked autoencoding baselines that adopt binary decomposition also degrade more on Class 4 (FreqMAE: 74.44/63.83; InfoMAE: 35.00/33.56). In contrast, standard contrastive and temporal self-supervised methods achieve higher scores on Class 4 (SimCLR: 95.06/95.19; TS-TCC: 89.11/90.53), even though these methods are not the strongest baselines on real-world datasets. TRIAD achieves the best performance across all classes (Class 1–Class 4). Overall, these results indicate that capturing the interaction-defined factor in Class 4 benefits from an explicit mechanism that allocates synergy to a dedicated component.

Table 7: SynTime per-class results (Acc % / F1 %). Class 1=Redundancy, Class 2=Uniqueness-A, Class 3=Uniqueness-B, Class 4=Synergy. Best in bold, second-best underlined.

Method	Class 1		Class 2		Class 3		Class 4		Overall	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
SimCLR	99.17	99.17	<u>99.89</u>	<u>99.83</u>	99.78	99.75	<u>95.06</u>	<u>95.19</u>	<u>98.48</u>	98.49
CMC	100.00	100.00	100.00	100.00	100.00	100.00	86.66	92.72	96.67	98.18
GMC	100.00	100.00	100.00	100.00	100.00	100.00	86.06	92.42	96.52	98.11
TNC	<u>99.89</u>	<u>99.92</u>	100.00	100.00	100.00	100.00	92.67	94.79	98.14	<u>98.68</u>
TS-TCC	97.94	96.92	99.44	99.33	99.33	98.98	89.11	90.53	96.46	96.44
TimeDRL	56.56	57.79	98.00	98.48	99.17	98.84	83.39	71.31	84.28	81.61
FOCAL	100.00	100.00	100.00	100.00	100.00	100.00	62.61	60.62	90.65	90.16
FreqMAE	98.39	99.19	100.00	99.23	100.00	<u>99.97</u>	74.44	63.83	93.21	90.56
InfoMAE	63.94	65.26	99.83	88.65	<u>99.83</u>	98.57	35.00	33.56	74.65	71.51
MATE	100.00	100.00	100.00	100.00	98.72	72.26	66.67	44.43	91.35	79.17
TRIAD	100.00	100.00	100.00	100.00	100.00	100.00	99.61	99.75	99.90	99.94

Table 8: Capacity-controlled ablation.

Dataset	TRIAD		No-PID	
	Acc	F1	Acc	F1
PAMAP2	93.62	91.86	82.90	77.91
HAR	92.95	92.87	87.89	87.64
WESAD	84.80	81.23	72.35	60.11
SDH	93.57	85.88	89.19	76.61

Table 9: Pretraining objective comparison.

Dataset	Reconstruction		Contrastive		TRIAD	
	Acc	F1	Acc	F1	Acc	F1
HAR	90.07	89.79	92.09	92.05	92.95	92.87
WESAD	76.39	71.18	73.20	66.56	84.80	81.23
SDH	89.65	77.63	89.20	77.41	93.57	85.88

B.2 Capacity-Controlled Ablation

To verify that the gains of TRIAD come from the PID-inspired objectives rather than the additional capacity of the routed architecture, we construct a *No-PID* variant that keeps the TRIAD backbone fixed and replaces all PID-inspired objectives with generic reconstruction losses. As shown in Table 8, TRIAD consistently and substantially outperforms this matched-capacity variant across all datasets, indicating that the predictive proxy objectives, rather than model size, drive the improvements.

B.3 Pretraining Objective Analysis

To evaluate whether the TRIAD architecture generalizes beyond the predictive proxy task, we replace the standard history-to-future prediction objective with two alternatives: masked reconstruction and contrastive learning. The routing architecture is kept unchanged, and only the objective-specific heads are modified.

Masked reconstruction masks a subset of patch tokens in the input sequence by zeroing the corresponding values and trains the decoders to reconstruct only the masked positions. Contrastive learning removes the reconstruction module, creates two masked views of the same sequence, and optimizes a SimCLR-style NT-Xent objective with a two-layer MLP projection head.

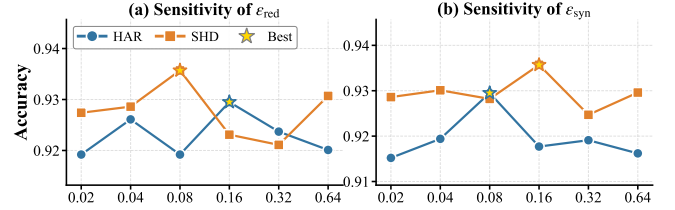
**Figure 10: Sensitivity analysis of margin thresholds.**

Table 9 shows that the standard TRIAD objective achieves the best performance on all three datasets. Masked reconstruction consistently outperforms contrastive learning on WESAD and SDH (e.g., WESAD Acc: 76.39 vs. 73.20), which suggests that token-level value recovery is more effective for physiological and building signals. On HAR, contrastive learning outperforms masked reconstruction (Acc: 92.09 vs. 90.07), which is consistent with stronger instance-level separability in motion patterns. Overall, the routing architecture supports multiple objectives, whereas history-to-future forecasting remains the strongest pretraining signal in this setting.

B.4 Sensitivity Analysis of Margin Thresholds

In this section, we analyze the sensitivity of our method to the margin threshold hyperparameters ϵ_{red} (redundancy margin) and ϵ_{syn} (synergy margin). These thresholds control the tolerance for identifying redundant and synergistic information components during the representation learning process.

As shown in Fig. 10(a), varying the redundancy margin ϵ_{red} yields stable performance on both datasets across the range $[0.02, 0.64]$, with accuracy variations within 1.5%. HAR achieves its best accuracy (92.95%) at $\epsilon_{red} = 0.16$, whereas SDH peaks (93.57%) at $\epsilon_{red} = 0.08$. Although the optimal settings differ across datasets, the overall stability suggests that redundancy extraction is robust to a wide choice of ϵ_{red} .

A similar trend holds for the synergy margin ϵ_{syn} (Fig. 10(b)). HAR reaches 92.95% at $\epsilon_{syn} = 0.08$, while SDH attains 93.57% at $\epsilon_{syn} = 0.16$, which is reversed relative to the best settings of ϵ_{red} . Across both datasets, the gap between the best and worst settings remains below 1.5%, indicating that TRIAD is insensitive to the precise margin choice and requires minimal hyperparameter tuning.