HRIL: Learning Multimodal Synergy via Higher-Order Tensor Modeling

Qun Dai[1], **Liangjian Wen**[1,2*], **Jiang Duan**[1,7], **Yong Dai**[3], **Dongkai Wang**[1], **Maolin Wang**[4], **Mingjie Wang**[5], **Jianzhuang Liu**[6], **HE YAN**[7], **Zhao Kang**[8]

[1]School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics
[2]Artificial Intelligence and Digital Finance Key Laboratory of Sichuan Province
[3]X-Humanoid
[4]Hong Kong Institute of AI for Science, City University of Hong Kong
[5]Zhejiang Sci-Tech University
[6]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[7]Chengdu Everimaging Science and Technology Co., Ltd
[8]University of Electronic Science and Technology of China
{daiqun1124,wlj6816}@gmail.com

Abstract

Self-supervised multimodal representation learning has achieved remarkable success across diverse domains, yet capturing *synergistic* information remains challenging due to the complexity of cross-modal interactions. Unlike the shared information across individual modalities, synergy arises when task-relevant signals emerge only from the joint configuration of multiple modalities and cannot be recovered from any modality in isolation. This work focuses on how to preserve the information capacity for such synergistic signals in multimodal representations. The key observation is that synergistic information is reflected in higher-order statistical dependence among modalities, which provides a principled target for explicitly modeling joint interactions. Motivated by this insight, we propose **H**igher-order **R**epresentation and **I**nformation **L**earning (HRIL), which constructs an empirical cross-moment tensor over modality embeddings to represent multi-way interactions. HRIL employs Tucker decomposition to obtain a core tensor, complemented by a synergy-aware regularizer that prevents energy concentration and preserves higher-order coupling capacity for synergistic information capture. Experiments on the controlled synergy task and real-world benchmarks demonstrate consistent improvements over existing multimodal contrastive methods, with notable gains on tasks dominated by synergistic interactions. Code is released at `https://github.com/brightest66/HRIL`.

1 Introduction

Self-supervised multimodal representation learning constructs joint representations from heterogeneous data sources whose underlying statistical structures differ fundamentally. Since distinct modalities arise from disparate data-generating processes and reside in qualitatively different statistical spaces, the central challenge is to leverage complementary and heterogeneous features across modalities for representations that are maximally informative for downstream tasks. Foundational pairwise contrastive methods such as CLIP [36] and ALIGN [18] have achieved remarkable transferability across vision, language, and audio domains, yet they operate under the *multiview redundancy*

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

assumption [41, 38, 48], which posits that task-relevant information is largely shared across modalities.

When we examine multimodal interactions across two or more modalities through the lens of Partial Information Decomposition (PID) [50, 4] — which decomposes task-relevant information into *redundancy* (shared across modalities), *uniqueness* (exclusive to individual modalities), and *synergy* (accessible only via joint observation of multiple modalities) — the limitations of redundancy-centric pairwise contrastive learning become apparent. For instance, in the MM-IMDb dataset [2], accurate genre prediction necessitates the synergistic integration of textual plot descriptions and visual poster elements, as the genre semantics emerge from their cross-modal combination rather than being fully encapsulated in either modality alone. Consequently, synergy is of fundamental importance in multimodal representation learning, as it embodies the irreducible cross-modal dependency that justifies multimodal representation learning and cannot be inferred from any single modality [11, 46].

A natural question is whether stronger pairwise contrastive learning can capture synergy. [11] demonstrates that cross-modal contrastive learning under the *multiview redundancy assumption* primarily captures shared information in bimodal settings. In this work, we further prove that a representation which is *sufficient* for all pairwise contrastive learning yet *minimal* in information content provably discards synergistic information accessible only through the joint observation of all modalities in scenarios with more than two modalities, because such information is not required for pairwise sufficiency, building on [43]. Several recent methods have moved beyond pairwise formulations. CoMM [11] allows synergy to emerge implicitly through augmented fusion, while Symile [37] and ConFu [23] target higher-order dependence via total correlation or subset fusion. However, these approaches either lack explicit guarantees for synergy preservation or model higher-order dependence without formally connecting it to task-relevant synergistic information.

To bridge this gap, we propose Higher-order Representation and Information Learning (HRIL), a synergy-aware fusion framework for multimodal contrastive representation learning. Guided by our analysis linking synergistic information to higher-order statistical dependence, HRIL models multi-way interactions explicitly by constructing an empirical cross-moment tensor over modality embeddings. HRIL then applies Tucker decomposition to obtain a compact interaction core and introduces a synergy-aware regularizer that prevents degenerate interaction structures and preserves diverse higher-order interaction patterns during training.

Our main contributions are threefold. **(1)** We formally establish the insufficiency of pairwise contrastive learning for synergy recovery among more than two modalities and identify the pure synergistic information as a conditional total-correlation difference, providing a precise information-theoretic criterion for when higher-order modeling is necessary. **(2)** We propose HRIL, which models higher-order interactions via empirical cross-moment tensorization and Tucker decomposition, coupled with a synergy-aware regularizer that penalizes excessive energy concentration in the Tucker core to preserve capacity for multi-way interactions. **(3)** We conduct extensive experiments on real-world benchmarks spanning healthcare and affective computing — which contain diverse modalities (*e.g.*, image, text, audio, *etc.*) — complemented by controlled synthetic evaluations. Our results demonstrate consistent improvements over state-of-the-art methods, with pronounced gains on synergy-heavy tasks.

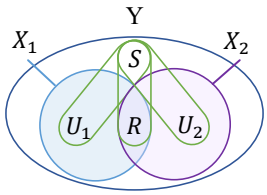

Figure 1: PID decomposes joint mutual information $I(X_1, X_2; Y)$ into shared R, unique U_1/U_2, and synergistic S.

2 Theoretical Analysis

In this section, we first formalize a latent-factor framework to demonstrate the insufficiency of pairwise contrastive learning for capturing synergy in scenarios involving more than two modalities, and then quantify the synergy gap via conditional total correlation, deriving an information-theoretic criterion for the necessity of higher-order modeling. In our theoretical analysis, we focus on the case of more than two modalities. Our analysis of pairwise contrastive learning focuses on its minimal sufficient representations. All proofs are provided in Appendix G.

2.1 A Latent-Factor Framework for Multimodal Representations

Task setup. Given $X = \{X_1, X_2, \ldots, X_M\}$ with M modalities and a downstream task label Y, our objective is to learn a multimodal representation $\mathbf{Z} = f(X)$ that is informative for Y. Throughout the paper, we use $X_{1:M} := (X_1, \ldots, X_M)$ and $X_A := \{X_i : i \in A\}$ for any index set $A \subseteq \{1, \ldots, M\}$.

Partial Information Decomposition (PID) [50, 4] conceptually decomposes task-relevant information into redundancy, uniqueness, and synergy (see Fig. 1 for a visual depiction). Motivated by this view, we adopt the latent-factor framework below.

Assumption 1 (Tripartite Latent Factor Decomposition)**.** *We assume the existence of latent factors $\{\mathbf{q}_r, \mathbf{q}_1, \ldots, \mathbf{q}_M, \mathbf{q}_{syn}\}$, and all task-relevant information is mediated by these latent factors, i.e.,*

$$Y \perp X_{1:M} \mid (\mathbf{q}_r, \mathbf{q}_1, \ldots, \mathbf{q}_M, \mathbf{q}_{syn}),$$

where $\mathbf{q}_r$ is shared factors across modalities, $\mathbf{q}_m$ denotes the m-th modality-specific factor, i.e., $X_{-m} \perp \mathbf{q}_m \mid (\mathbf{q}_r, \mathbf{q}_{-m}, \mathbf{q}_{\text{syn}})$, where $X_{-m} = X_{\{1,\ldots,M\}\setminus\{m\}}$. And $\mathbf{q}_{syn}$ denotes synergistic factors that capture task-relevant information emerging only from joint observation. These latent factors are mutually independent. Throughout, $\perp$ denotes conditional independence.

Based on the above, we further state that $(\mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}})$ represents a decomposition of the task-relevant information contained in the complete multimodal observation $X_{1:M}$. Accordingly, once $X_{1:M}$ is given, $(\mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}})$ provides no further information about Y, which is expressed as

$$Y \perp (\mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}}) \mid X_{1:M}.$$

Among these latent factors, $\mathbf{q}_{\text{syn}}$ is substantially harder to capture because it corresponds to task-relevant dependencies that emerge only through joint multimodal observation and therefore requires a more precise characterization [11, 46]. We formalize this accessibility constraint as follows.

Definition 1 (Pure Synergistic Factor)**.** *We call $\mathbf{q}_{syn}$ a pure synergistic factor relative to $\mathbf{q}_r$ if*

$$I(\mathbf{q}_{syn}; X_{1:M} \mid \mathbf{q}_r) > 0, \quad \textit{and} \quad I(\mathbf{q}_{syn}; X_A \mid \mathbf{q}_r) = 0, \ \forall\, A \subsetneq \{1, \ldots, M\}.$$

Under Assumption 1, the observations are generated as $X_{1:M} \sim p(X_{1:M} \mid \mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}})$, where each unique factor $\mathbf{q}_m$ influences only X_m. Motivated by this latent-factor framework, we interpret the learned multimodal representation $\mathbf{Z}$ as being organized into three functional subspaces:

$$\mathbf{Z} \approx V^r f_r(\mathbf{q}_r) + V^u f_u(\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_M) + V^{\text{syn}} f_{\text{syn}}(\mathbf{q}_{\text{syn}}) + \epsilon, \tag{1}$$

where V^r spans a shared subspace, V^u captures modality-specific structures, V^{syn} denotes a synergistic subspace, and ϵ is a noise term.

2.2 Minimal Sufficient Representation Analysis

Let $\mathbf{Z}^{\min}_{\text{pair}}$ denote a representation that is *sufficient* for all pairwise inter-modal alignments and *minimal* in the sense of minimizing $I(\mathbf{Z}; X_{1:M})$ over the class of such representations. Inspired by [43], we adopt the following shared information characterization for our analysis:

$$I(\mathbf{Z}^{\min}_{\text{pair}}; X_{1:M} \mid \mathbf{q}_r) = 0, \tag{2}$$

meaning that $\mathbf{Z}^{\min}_{\text{pair}}$ retains only information captured by the shared factor $\mathbf{q}_r$.

Prior work [40, 43] has shown that minimal sufficient representations in pairwise contrastive learning can be inadequate for downstream tasks. Here, we sharpen this limitation within this latent-factor framework (Assumption 1), isolating the synergistic component excluded by pairwise minimal sufficiency.

Theorem 1 (Insufficiency of Pairwise Minimal Sufficient Representation)**.** *Under Assumption 1, suppose that $\mathbf{q}_{syn}$ is a pure synergistic factor in the sense of Definition 1 and that $I(Y; \mathbf{q}_{syn} \mid \mathbf{q}_r, \mathbf{q}_{1:M}) > 0$. Then*

$$I(Y; \mathbf{Z}^{\min}_{pair}) \leq I(Y; \mathbf{q}_r).$$

Moreover, the total information gap is lower bounded as

$$I(Y; X_{1:M}) - I(Y; \mathbf{Z}^{\min}_{pair}) \ \geq \ \underbrace{I(Y; \mathbf{q}_{1:M} \mid \mathbf{q}_r)}_{\textit{unique gap}} + \underbrace{I(Y; \mathbf{q}_{syn} \mid \mathbf{q}_r, \mathbf{q}_{1:M})}_{\textit{synergistic gap}}.$$

Theorem 1 reveals that pairwise contrastive learning even after recovers all shared and unique factors $(\mathbf{q}_r, \mathbf{q}_{1:M})$, an irreducible synergistic gap $I(Y; \mathbf{q}_{\text{syn}} \mid \mathbf{q}_r, \mathbf{q}_{1:M}) > 0$ persists and can only be closed by information about $\mathbf{q}_{\text{syn}}$. Inspired by [23], we turn to total correlation [45] as a measure of higher-order dependence to characterize pure synergistic information. Specifically, total correlation characterizes the overall statistical coupling among multiple random variables by the Kullback–Leibler divergence between their joint distribution and the product of their marginals. The conditional difference below makes its connection to pure synergy explicit.

Proposition 1 (Conditional Total Correlation Decomposition)**.** *For M modalities, the conditional total correlation admits the chain-rule decomposition*

$$\mathrm{TC}(X_{1:M} \mid \mathbf{q}_r) = \sum_{k=2}^{M} I(X_k;\, X_1, \ldots, X_{k-1} \mid \mathbf{q}_r).$$

In particular, for $M = 3$: $\mathrm{TC}(X_1, X_2, X_3 \mid \mathbf{q}_r) = I(X_1; X_2 \mid \mathbf{q}_r) + I(X_3; X_1, X_2 \mid \mathbf{q}_r)$, and symmetrically for all permutations.

Building on this decomposition, we now establish the precise quantitative link between synergistic information and higher-order statistical dependence.

Theorem 2 (Synergy Total Correlation Identity)**.** *Under Definition 1, the synergistic information satisfies the identity*

$$I(\mathbf{q}_{syn}; X_{1:M} \mid \mathbf{q}_r) = \mathrm{TC}(X_{1:M} \mid \mathbf{q}_r, \mathbf{q}_{syn}) - \mathrm{TC}(X_{1:M} \mid \mathbf{q}_r).$$

Interpretation. We unpack Theorem 2 in the following steps. **(1) Synergy changes joint dependence without altering single-modality marginals.** By Definition 1, conditioning on $\mathbf{q}_{\text{syn}}$ leaves each marginal entropy $H(X_m \mid \mathbf{q}_r)$ unchanged, but reduces the joint entropy $H(X_{1:M} \mid \mathbf{q}_r)$. Since $\mathrm{TC} = \sum_m H(X_m \mid \cdot) - H(X_{1:M} \mid \cdot)$, this reduction increases TC exactly by $I(\mathbf{q}_{\text{syn}}; X_{1:M} \mid \mathbf{q}_r)$. **(2) The resulting dependence is higher-order.** Conditional on $\mathbf{q}_{syn}$ also preserves every pairwise marginal distribution given $\mathbf{q}_r$; hence the TC difference in Theorem 2 can only be attributed to irreducible multi-way dependence, which pairwise objectives cannot access.

3 Method

The overall pipeline of HRIL is illustrated in Fig. 2. Let $\mathbf{h}_v^{(m)} = f_m(X_v^{(m)})$ denote the unimodal encoder features, a shared Transformer g produces $\mathbf{Z}_{v,u}^{(m)} = g(\mathbf{h}_v^{(m)})$ and $\mathbf{Z}_{v,f} = g(\mathbf{h}_v^{(1)}, \ldots, \mathbf{h}_v^{(M)})$, where $v \in \{1, 2\}$ indexes the augmented views. Given the encoded modality embeddings, HRIL optimizes three complementary components. First, a modality-fusion contrastive objective aligns each unimodal representation with the fused representation to extract cross-modal task-relevant signals. Second, the synergy-aware regularization preserves the capacity of the fused representation to encode higher-order synergistic interactions. Third, an auxiliary cross-view alignment objective stabilizes the shared and modality-specific subspaces while preventing representational collapse.

3.1 Higher-Order Dependence Modeling for the Synergistic Subspace

Theorems 1 and 2 isolate purely cross-modal, higher-order dependence as the sole information-theoretic carrier of synergy. To translate this criterion into a practical objective, we require a mechanism to distinguish genuine multi-way interactions in the embedding space. Inspired by [1, 21, 53], we employ the *empirical cross-moment tensor* that aggregates cross-modal dependencies across all interaction orders. Here $\otimes$ denotes the vector outer product and $\times_m$ denotes multiplication along tensor mode m. For a fixed augmented view, given a mini-batch of B samples with modality embeddings $\mathbf{Z}_m^{(b)} \in \mathbb{R}^{d_m}$, the M-th order empirical cross-moment tensor $\mathcal{T}$ is defined as

$$\mathcal{T} := \frac{1}{B} \sum_{b=1}^{B} \mathbf{Z}_1^{(b)} \otimes \mathbf{Z}_2^{(b)} \otimes \cdots \otimes \mathbf{Z}_M^{(b)} \;\in\; \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_M}, \tag{3}$$

whose $(i_1, \ldots, i_M)$-th entry estimates the mixed moment $\mathbb{E}[Z_{1,i_1} \cdots Z_{M,i_M}]$. Its Tucker decomposition [21] $\mathcal{T} \approx \mathcal{G} \times_1 \mathbf{A}_1 \cdots \times_M \mathbf{A}_M$, where $\mathbf{A}_m \in \mathbb{R}^{d_m \times r_m}$ are factor matrices, isolates the

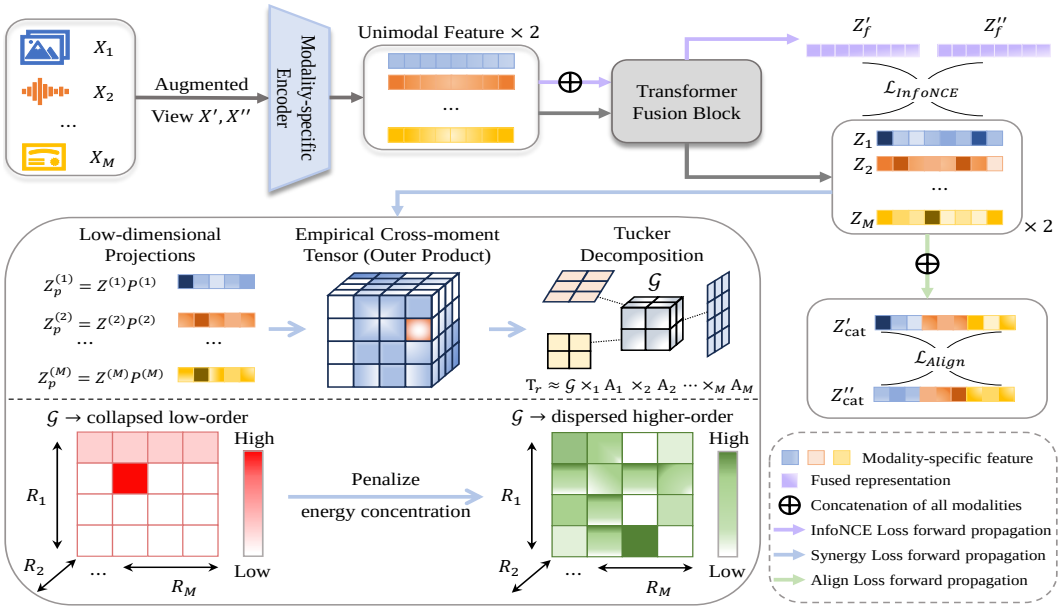

Figure 2: The overall pipeline of HRIL. Given M modalities, we apply data augmentation to construct two views X' and X'', and encode them into unimodal representations $\{\mathbf{Z}'_m\}_{m=1}^M$ and $\{\mathbf{Z}''_m\}_{m=1}^M$, as well as fused representations $\mathbf{Z}'_f$ and $\mathbf{Z}''_f$. The contrastive objective $\mathcal{L}_{\text{InfoNCE}}$ performs cross-view alignment between each unimodal representation and the fused one, while $\mathcal{L}_{\text{syn}}$ and $\mathcal{L}_{\text{align}}$ are computed from unimodal representations in both views.

core tensor $\mathcal{G}$ whose *multilinear rank* characterizes the multilinear complexity of the modeled cross-moment tensor. If the modalities interact only through a low-dimensional shared factor, $\mathcal{T}$ inherits a low multilinear rank and its energy concentrates into a few separable components; preserving capacity for higher-order dependencies motivates discouraging excessive energy concentration in $\mathcal{G}$. This suggests a natural regularization strategy: **penalize energy concentration in $\mathcal{G}$ to preserve representational capacity for higher-order interactions, and thus for synergy**. The following definition and lemmas formalize this diagnostic.

Definition 2 (Redundant Low-Order Structure)**.** *Multimodal embeddings $\{\mathbf{Z}_i\}_{i=1}^M$ are said to exhibit a* redundant low-order structure *if $\mathbf{Z}_i \approx \sum_{j\neq i} \mathbf{W}_{ij}\mathbf{Z}_j$ with $\operatorname{rank}(\mathbf{W}_{ij}) = r_{ij} \ll d_j$, where the approximation holds in the population sense: the reconstruction residual $\boldsymbol{\delta}_i := \mathbf{Z}_i - \sum_{j\neq i} \mathbf{W}_{ij}\mathbf{Z}_j$ satisfies $\mathbb{E}[\|\boldsymbol{\delta}_i\|^2] \to 0$ under the shared-factor dominance regime.*

Under Definition 2, when the shared component dominates the embedding energy, the low-rank predictor $\mathbf{W}_{ij}$ explains nearly all of $\mathbf{Z}_i$, and the embeddings collapse into the redundant low-order regime. We formally state this result in Proposition 2. The following two lemmas connect redundant low-order structure to multilinear complexity and characterize single-component degeneration of the interaction core.

Lemma 1 (Redundancy Implies Low Multilinear Complexity)**.** *If the embeddings exhibit a redundant low-order structure in the sense of Definition 2, then the cross-moment tensor $\mathcal{T} = \mathbb{E}[\bigotimes_m \mathbf{Z}_m]$ admits a low multilinear-complexity approximation. In particular, the Tucker core $\mathcal{G}$ of $\mathcal{T}$ has at least one mode whose effective multilinear rank satisfies:*

$$\operatorname{rank}_{(k)}(\mathcal{G}) \leq \sum_{j\neq i} r_{ij} \quad \textit{for modality } k = i.$$

Lemma 2 (Energy Concentration Indicates Degenerate Structure)**.** *Let*

$$\rho(\mathcal{G}) := \max_{i_1,\ldots,i_M} \frac{\mathcal{G}^2_{i_1,\ldots,i_M}}{\|\mathcal{G}\|_F^2}.$$

If $\rho(\mathcal{G}) \geq 1 - \delta$ for some $\delta \in [0, 1]$, then $\mathcal{G}$ is within relative Frobenius error at most $\sqrt{\delta}$ of a one-sparse core tensor. For small δ, this indicates dominance by a single separable component in the Tucker core.

Definition 2, Lemma 1, and Lemma 2 connect information theory and representation geometry: when shared factors dominate the embeddings, the cross-modal predictor is low-rank, which constrains the multilinear rank of the core tensor $\mathcal{G}$; extreme energy concentration indicates dominance by a single separable component. Penalizing this concentration therefore preserves the representational capacity needed to encode the higher-order dependence that, by Theorem 2, carries pure synergistic information.

To avoid the cost of computing the full-dimensional cross-moment tensor, we first project each embedding into an r_p-dimensional lower-dimensional subspace. For each augmented view $v \in \{1, 2\}$, we define $\mathbf{Z}_{p,v}^{(m,b)} = \mathbf{Z}_{v,u}^{(m,b)} \mathbf{P}^{(m)}$, where $\mathbf{P}^{(m)} \in \mathbb{R}^{d_m \times r_p}$ is constructed by QR decomposition of a random Gaussian matrix with independent entries and $r_p < \min_m d_m$. This projection preserves pairwise geometry up to a multiplicative distortion factor by the Johnson–Lindenstrauss lemma (Lemma 3) [10]. We then compute the projected empirical cross-moment tensor for each view over the mini-batch:

$$\mathcal{T}_{r_p,v} := \frac{1}{B} \sum_{b=1}^{B} \mathbf{Z}_{p,v}^{(1,b)} \otimes \mathbf{Z}_{p,v}^{(2,b)} \otimes \cdots \otimes \mathbf{Z}_{p,v}^{(M,b)} \in \mathbb{R}^{r_p \times r_p \times \cdots \times r_p}, \tag{4}$$

and perform Tucker decomposition $\mathcal{T}_{r_p,v} \approx \mathcal{G}_v \times_1 \mathbf{A}_{1,v} \cdots \times_M \mathbf{A}_{M,v}$, yielding the core tensor $\mathcal{G}_v$.

To prevent excessive core-energy concentration, we introduce the following margin-based synergy-aware regularization grounded in Lemma 2:

$$\mathcal{L}_{\text{syn}} = \frac{1}{2} \sum_{v=1}^{2} \left[\xi - (1 - \rho(\mathcal{G}_v)) \right]^+, \tag{5}$$

where $\xi \in (0, 1)$ is a margin hyperparameter controlling the minimum energy dispersion. Throughout this paper, we adopt the notation $[\cdot]^+ = \text{ReLU}(\cdot)$. By penalizing $\rho(\mathcal{G}_v) > 1 - \xi$ in each view, $\mathcal{L}_{\text{syn}}$ discourages one-sparse degeneration in the Tucker cores, thereby preserving multi-way coupling capacity.

While $\mathcal{L}_{\text{syn}}$ preserves the capacity for higher-order dependencies, it does not actively recover synergistic information. To learn such information, we employ a contrastive alignment strategy from [11], which aligns each unimodal representation with the fused multimodal representation, *i.e.*,

$$\mathcal{L}_{\text{InfoNCE}}^{(m)} := -\frac{1}{2} \left(\hat{I}_{\text{NCE}}(\mathbf{Z}_{1,u}^{(m)}, \mathbf{Z}_{2,f}) + \hat{I}_{\text{NCE}}(\mathbf{Z}_{2,u}^{(m)}, \mathbf{Z}_{1,f}) \right). \tag{6}$$

Here $\hat{I}_{\text{NCE}}$ is the InfoNCE mutual-information estimator [33]:

$$\hat{I}_{\text{NCE}}(\mathbf{Z}, \mathbf{Z}') = \mathbb{E}_{\substack{\mathbf{z}, \mathbf{z}'_{\text{pos}} \sim p(\mathbf{Z}, \mathbf{Z}') \\ \mathbf{z}'_{\text{neg}} \sim p(\mathbf{Z}')}} \left[\log \frac{\exp \text{sim}(\mathbf{z}, \mathbf{z}'_{\text{pos}})}{\exp \text{sim}(\mathbf{z}, \mathbf{z}'_{\text{pos}}) + \sum_{\mathbf{z}'_{\text{neg}}} \exp \text{sim}(\mathbf{z}, \mathbf{z}'_{\text{neg}})} \right].$$

We use $\text{sim}(\cdot, \cdot)$ to denote cosine similarity throughout this work. $\mathbf{Z}_{1,u}^{(m)}$ and $\mathbf{Z}_{2,u}^{(m)}$ denote the m-th unimodal embeddings from the two views, and $\mathbf{Z}_{1,f}$, $\mathbf{Z}_{2,f}$ denote the corresponding fused representations. Critically, although the above loss objective only captures shared and unique information in [11], the proposed $\mathcal{L}_{\text{syn}}$ discourages excessive energy concentration within the core tensor to maintain the higher-order representational capacity of $\mathbf{Z}_f$. Consequently, the identical contrastive loss further enables the model to extract synergistic components.

3.2 Contrastive alignment for shared and unique subspaces

Although Eq. (5) preserves capacity for synergistic interactions, it does not by itself control the geometry of unimodal embeddings. This is important because multimodal InfoNCE can suffer from competing alignment directions and attraction–repulsion conflicts, leading to modality gaps and unstable optimization [51]. Inspired by [3, 56], we introduce an auxiliary alignment regularizer that encourages cross-view consistency while preserving variation in modality-specific directions.

For each view $v \in \{1, 2\}$, let $\hat{\mathbf{Z}}_i \in \mathbb{R}^{B \times D}$ denote the ℓ_2-normalized representation that concatenates all unimodal embeddings along the feature dimension for the i-th view, where B is the batch size and D is the total feature dimension. The alignment objective is:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{var}}, \tag{7}$$

where

$$\mathcal{L}_{\text{inv}} = \frac{1}{B} \sum_{b=1}^{B} \big(1 - \text{sim}(\hat{\mathbf{Z}}_1^{(b)}, \hat{\mathbf{Z}}_2^{(b)})\big)^2 \quad \text{and} \quad \mathcal{L}_{\text{var}} = \frac{1}{D} \sum_{d=1}^{D} \left[\gamma - \sqrt{\text{Var}\big(\bar{\mathbf{Z}}_d\big) + \epsilon}\right]^+,$$

with $\bar{\mathbf{Z}} = (\hat{\mathbf{Z}}_1 + \hat{\mathbf{Z}}_2)/2$, where γ is the variance threshold and ϵ is a small constant.

Among these two terms, $\mathcal{L}_{\text{inv}}$ encourages cross-view agreement of shared factors at the sample level, and $\mathcal{L}_{\text{var}}$ penalizes standard deviations below γ (0.5 by default) along each embedding dimension to discourage coordinate-wise collapse. Because each modality embedding $\mathbf{Z}_v^{(m)}$ is individually normalized before concatenation, modality-specific directions are preserved in $\hat{\mathbf{Z}}_v$; $\mathcal{L}_{\text{var}}$ then encourages variance along *all* dimensions, helping preserve modality-specific variation during alignment.

Overall objective. The final loss combines all three components:

$$\mathcal{L}_{\text{total}} = \sum_{m=1}^{M} \mathcal{L}_{\text{InfoNCE}}^{(m)} + \lambda_a \mathcal{L}_{\text{align}} + \lambda_s \mathcal{L}_{\text{syn}}, \tag{8}$$

where λ_a and λ_s are hyperparameters.

4 Experiments

We evaluate HRIL on both synthetic and real-world benchmarks to validate its effectiveness in capturing synergistic interactions and enhancing downstream task performance. In the synthetic setting, we follow the Trifeature dataset protocol from [11], which provides a controlled environment to explicitly isolate redundancy, uniqueness, and synergy. To assess practical utility, we conduct extensive experiments on Multibench [27], a comprehensive multimodal benchmark spanning diverse domains and modality combinations (*e.g.*, healthcare, affective computing, *etc.*). First, we evaluate the trimodal setting, which naturally exhibits higher-order statistical dependencies essential for synergy, thereby directly testing our method's capacity to model complex multi-way interactions. Subsequently, we evaluate generalization in bimodal configurations, which constitute the predominant configuration in practical applications. For evaluation, we use linear probing, *i.e.*, freezing the pre-trained feature extractor and training a linear classifier on top of the learned representations. The resulting downstream performance serves as a direct proxy for the discriminative quality and generalization capacity of the learned multimodal representations. Detailed experimental settings are provided in Appendix B.

4.1 Experiments on Synthetic Datasets with Trifeature

We follow the Trifeature [17] synthetic protocol of CoMM [11]. For redundancy and uniqueness, *shape* is treated as shared and *texture* as modality-specific. The dataset consists of image pairs sharing identical shapes but exhibiting independent textures. We formulate two linear probing objectives: (1) predicting the common shape class (redundancy), and (2) identifying the texture from a single modality (uniqueness), both yielding a 10% random-guessing baseline. To evaluate synergy, we impose a deterministic cross-modal mapping $\mathcal{M}$ that links the texture of X_1 to the color of X_2 (*e.g.*, stripes $\mapsto$ red), restricting training exclusively to pairs that adhere to this constraint.

Table 1: Linear probing accuracy (in %) for redundancy (shape), uniqueness (texture), and synergy (color and texture) on the Trifeature dataset. ♣ denotes results from [11].

Model	*Redundancy*↑	*Uniqueness*↑	*Synergy*↑
Cross♣ [36]	**100.0**	11.6	50.0
Cross+Self♣ [52]	99.7	86.9	50.0
FactorCL♣ [26]	99.8	62.5	46.5
CoMM [11]	$99.9_{\pm 0.06}$	$86.8_{\pm 2.99}$	$71.4_{\pm 3.47}$
HRIL (ours)	$99.7_{\pm 0.14}$	$\mathbf{92.6_{\pm 3.04}}$	$\mathbf{82.3_{\pm 1.61}}$

Evaluation proceeds via a binary compliance task defined as $Y = \mathbb{1}[\,(\mathrm{tex}(X_1), \mathrm{col}(X_2)) \in \mathcal{M}\,]$, with a 50% chance baseline.

Experimental results are summarized in Tab. 1. Cross-modality constraints with InfoNCE [36] ("Cross") effectively capture redundant shape information but yield near-chance performance on uniqueness. Self-supervised constraints on each encoder ("Cross+Self"), FactorCL [26], and CoMM [11] all improve uniqueness performance, with CoMM achieving the best result as it implicitly models the synergistic factor and is the only baseline that significantly improves synergy capture. In comparison, HRIL achieves the highest uniqueness and synergy performance, outperforming CoMM by 5.8% on uniqueness and 10.9% on synergy.

4.2 Experiments on Multibench

To further validate the practical utility of HRIL, we conduct extensive experiments on Multibench [27], which includes multiple datasets with different modalities and domains. These datasets provide a comprehensive evaluation of the model's ability to learn multimodal representations that are effective for downstream tasks. Further dataset details are provided in Appendix C.

4.2.1 Experiments with 3 Modalities: UR-FUNNY, MOSI, and MOSEI

We first evaluate the trimodal setting since it naturally exhibits higher-order statistical dependencies essential for synergy. The data preprocessing steps can be found in Appendix C. Those experiments are conducted in trimodal configurations using identical encoders and modality arrangements, with all models trained directly on encoded representations. We compare against several established multimodal contrastive learning frameworks: Tri_CLIP (trained with pairwise contrastive objectives across three modalities), ConFu [23], CMC [39], and CoMM [11]. As shown in Tab. 2, while ConFu extends Tri_CLIP by incorporating a higher-order contrastive loss, it yields only limited performance gains. In contrast, HRIL consistently surpasses all baselines across the three datasets and outperforms the strongest competitor — CoMM — by 0.7% on UR-FUNNY, 1.2% on MOSI, and 0.6% on MOSEI. These empirical results demonstrate that HRIL's explicit modeling of higher-order interactions effectively enhances representation quality and downstream task performance, thereby validating the theoretical principles established in our information-theoretic analysis.

Table 2: Linear probing top-1 accuracy (in %) for classification tasks on trimodal Multibench.

Model	*UR-FUNNY*↑	*MOSI*↑	*MOSEI*↑	**Average**
Tri_CLIP	$60.6_{\pm 0.79}$	$60.1_{\pm 3.43}$	$64.1_{\pm 0.42}$	61.6
CMC [39]	$60.1_{\pm 0.78}$	$62.4_{\pm 1.67}$	$62.7_{\pm 0.72}$	61.73
ConFu [23]	$61.3_{\pm 0.96}$	$62.5_{\pm 2.66}$	$64.8_{\pm 0.61}$	62.87
CoMM [11]	$64.8_{\pm 1.13}$	$66.0_{\pm 2.27}$	$69.9_{\pm 0.34}$	66.9
HRIL (ours)	$\mathbf{65.5_{\pm 0.56}}$	$\mathbf{67.2_{\pm 1.28}}$	$\mathbf{70.5_{\pm 0.20}}$	**67.73**

4.2.2 Experiments with 2 Modalities

Beyond the trimodal setting, we further evaluate whether HRIL remains effective in the more common bimodal scenario and test the practical robustness of our synergy-aware objective across heterogeneous domains, including healthcare and affective understanding. Adopting the data preprocessing steps established in [26], we compare our method with Cross [36], Cross+Self [52], FactorCL [26], and CoMM [11]. Tab. 3 summarizes the results. HRIL achieves the best performance on all four bimodal classification tasks. Compared with the strongest baseline, CoMM, HRIL improves accuracy by 1.6 percentage points on MIMIC, 2.1 on MOSI, 0.6 on UR-FUNNY, and 2.2 on MUStARD, leading

Table 3: Linear probing top-1 accuracy (in %) for classification tasks on bimodal Multibench. ♣ denotes results are from [26].

Model	*Classification*				
	MIMIC↑	*MOSI*↑	*UR-FUNNY*↑	*MUStARD*↑	**Average**↑
Cross♣ [36]	$66.7_{\pm 0.1}$	$47.8_{\pm 1.8}$	$50.1_{\pm 1.9}$	$53.5_{\pm 2.9}$	54.52
Cross+Self♣ [52]	$65.49_{\pm 0.0}$	$49.0_{\pm 1.1}$	$59.9_{\pm 0.9}$	$53.9_{\pm 4.0}$	57.07
FactorCL♣ [26]	$67.3_{\pm 0.0}$	$51.2_{\pm 1.6}$	$60.5_{\pm 0.8}$	$55.80_{\pm 0.9}$	58.7
CoMM [11]	$66.4_{\pm 0.4}$	$63.7_{\pm 2.5}$	$63.3_{\pm 0.5}$	$64.4_{\pm 1.1}$	64.45
HRIL (ours)	$\mathbf{68.0_{\pm 0.7}}$	$\mathbf{65.8_{\pm 1.2}}$	$\mathbf{63.9_{\pm 0.9}}$	$\mathbf{66.6_{\pm 2.7}}$	**66.08**

to an average gain of 1.63 percentage points. These results indicate that the proposed synergy-aware regularization is not limited to interactions among three or more modalities; it also improves representation learning in practical bimodal settings by preserving non-degenerate cross-modal dependence.

5 Ablation Studies

Loss functions. To evaluate the contribution of each loss term in HRIL, we compare different loss combinations on the Trifeature dataset for capturing multimodal interactions.

Table 4: Linear probing accuracy (in %) of multimodal interactions on the Trifeature dataset under different combinations of loss terms in Eq. 8.

Loss components			*Redundancy*↑	*Uniqueness*↑	*Synergy*↑
$\mathcal{L}_{\text{InfoNCE}}$	$\mathcal{L}_{\text{align}}$	$\mathcal{L}_{\text{syn}}$			
×	✓	✓	$28.5_{\pm 2.74}$	$12.5_{\pm 3.15}$	$50.0_{\pm 0.00}$
✓	×	×	$99.9_{\pm 0.10}$	$87.2_{\pm 2.13}$	$50.0_{\pm 0.00}$
✓	×	✓	$99.9_{\pm 0.05}$	$86.6_{\pm 2.79}$	$72.4_{\pm 3.16}$
✓	✓	×	$99.9_{\pm 0.09}$	$85.1_{\pm 6.86}$	$78.0_{\pm 0.44}$
✓	✓	✓	$99.7_{\pm 0.14}$	$\mathbf{92.6_{\pm 3.04}}$	$\mathbf{82.3_{\pm 1.61}}$

As shown in Tab. 4, both $\mathcal{L}_{\text{InfoNCE}}$ alone and $\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{syn}}$ perform poorly on synergy, achieving only chance-level accuracy (50.0%). The variant without $\mathcal{L}_{\text{InfoNCE}}$ performs worse overall, with substantially lower redundancy and uniqueness accuracy. Adding $\mathcal{L}_{\text{syn}}$ to $\mathcal{L}_{\text{InfoNCE}}$ increases synergy accuracy to 72.4%, while adding $\mathcal{L}_{\text{align}}$ increases it to 78.0%. The full objective further improves synergy accuracy from 78.0% to 82.3% and uniqueness accuracy from 85.1% to 92.6% compared with $\mathcal{L}_{\text{InfoNCE}} + \mathcal{L}_{\text{align}}$. Meanwhile, redundancy remains nearly saturated. These results suggest that HRIL benefits from the complementary contributions of all components.

Interaction attribution. To evaluate whether the proposed synergy-aware regularization (Eq. (5)) increases cross-modal interaction attribution, we report InterSHAP [49] as an attribution-based interaction metric. Tab. 5 reports InterSHAP values on three trimodal datasets from Multibench [27]. Without $\mathcal{L}_{\text{syn}}$ (Eq. (5)), the interaction attribution remains modest. In contrast, adding the proposed synergy-aware regularization consistently raises mean InterSHAP on all datasets, improving the average score from 5.32% to 6.87%. The gains are most pronounced on MOSI and UR-FUNNY, indicating stronger use of cross-modal interaction patterns after applying the synergy-aware regularization. The slight increase on MOSEI indicates a smaller mean change in interaction attribution. Moreover, since interaction attributions in real-world benchmarks are often numerically small [49], the consistent upward shift across datasets provides complementary behavioral evidence that $\mathcal{L}_{\text{syn}}$ encourages the model to exploit cross-modal interactions. The metric definition and further comparisons with CoMM and a multilinear-rank control are provided in Appendix D.1.

Table 5: Quantitative evaluation of InterSHAP on Multibench [27] datasets. × denotes the model without $\mathcal{L}_{\text{syn}}$, and ✓ denotes the full objective of HRIL.

$\mathcal{L}_{\text{syn}}$	*InterSHAP Value (%)*			**Average**
	UR-FUNNY	*MOSI*	*MOSEI*	
×	$7.78_{\pm 4.56}$	$6.82_{\pm 3.73}$	$1.35_{\pm 0.53}$	5.32
✓	$9.25_{\pm 5.56}$	$9.90_{\pm 4.81}$	$1.47_{\pm 0.40}$	6.87
Improvement	$\mathbf{+1.47}$	$\mathbf{+3.08}$	$\mathbf{+0.12}$	$\mathbf{+1.55}$

Fusion module. We compare the attention-based fusion module with a linear fusion baseline on the Trifeature dataset. As shown in Tab. 6, the linear fusion module fails to capture synergistic information, whereas HRIL substantially improves the performance of synergy. This result suggests that synergy-aware regularization is effective only when the fusion module has sufficient capacity to model higher-order cross-modal interactions. It also provides empirical

Table 6: Linear probing accuracy (in %) of multimodal interactions on the Trifeature dataset. Concat+Linear: concatenate the modality embeddings and pass through a linear layer.

Fusion	R	U	S	**Average**
Concat+Linear	$99.9_{\pm 0.1}$	$65.4_{\pm 10.73}$	$50.0_{\pm 0.00}$	71.77
HRIL	$99.7_{\pm 0.14}$	$\mathbf{92.6_{\pm 3.04}}$	$\mathbf{82.3_{\pm 1.61}}$	**91.83**

support for Definition 2: if the information in one modality can be linearly approximated from another modality, the resulting representation is unlikely to encode genuinely synergistic information.

6 Related Work

Multimodal contrastive learning. Multimodal contrastive learning aligns cross-modal representations by maximizing mutual information through constructed positive and negative pairs [36, 33, 39]. Positive pairs are typically constructed by augmenting the same instance, leveraging intra- and cross-modal variations as mutual supervision signals [19, 5, 7, 56]. The effective utilization of the three forms of multimodal information (*i.e.*, redundancy, uniqueness, and synergy) is driving the advancement of representation learning [47]. Predominant pairwise contrastive frameworks primarily capture shared information across modalities [36, 18], often at the expense of modality-unique features. MV-DHEL [22] jointly aligns all views of the same sample while imposing view-specific uniformity to decouple alignment from uniformity. While recent efforts have pivoted toward disentangling and preserving modality-specific information [26, 44, 15, 35], and methods such as CoMM [11] and InfMasking [46] implicitly capture synergistic interactions, the explicit modeling and recovery of synergy remain an open challenge. Building on CoMM's fusion and alignment paradigm, HRIL adds an explicit higher-order structural regularizer.

Higher-order multimodal dependence. Recent self-supervised methods move beyond pairwise alignment by explicitly modeling multi-way cross-modal dependence. Symile [37] uses a scalar multilinear critic to optimize a contrastive lower bound on total correlation, while ConFu [23] aligns fused modality subsets with the remaining modality alongside pairwise alignment. M3G [34] promotes joint consistency across views by minimizing the gap between ground-truth and optimal matching costs under entropy-regularized multi-marginal optimal transport. Geometric criteria provide an alternative route to joint alignment: GRAM [9] aligns modalities by minimizing the Gramian volume spanned by their embeddings, and TRIANGLE [8] enforces tri-modal alignment by replacing cosine similarity with a triangle-area similarity computed in the embedding span. Nevertheless, these approaches largely target joint dependence or alignment, while HRIL additionally regularizes the empirical cross-moment structure to preserve capacity for synergistic interactions.

Tensor-based multimodal fusion. Tensor fusion provides an explicit mechanism for modeling multiplicative cross-modal interactions [1, 21]. TFN [53] enumerates unimodal, bimodal, and trimodal interactions by forming outer products of modality embeddings, but incurs exponential complexity. LMF [28] alleviates this via low-rank factorization while retaining multiplicative coupling. Diverging from tensor-based fusion paradigms, we instead employ Tucker decomposition [21] as a *regularization* prior, explicitly enforcing a non-degenerate multi-way structure in the fused representation.

7 Conclusion

We presented HRIL, a multimodal contrastive representation learning framework that addresses the synergistic gap left by pairwise contrastive learning. By proving under the condition of pure synergy that synergistic information equals a conditional total-correlation difference recoverable only through higher-order dependence, we motivated synergy-aware regularization on the Tucker-decomposed cross-moment core that preserves multi-way interaction capacity alongside multimodal contrastive objectives. Experiments on synthetic and real-world benchmarks show consistent gains, especially on the controlled synergy task. Current limitations include the assumption of cleanly separated latent factors; future work will explore relaxing this assumption and extending the framework to large multimodal models. Further discussions can be found in Appendix A.

Acknowledgments This work was supported by the Sichuan Provincial Innovation Group Project (Grant No. 2024NSFTD0054), the Major Science and Technology Special Project of the Sichuan Provincial Department of Science and Technology (Grant No. 2024ZDZX0002), Fundamental Research Funds for the Central Universities (JBK202511081), the National Natural Science Foundation of China (No. U24A20323 , No. 62502397, No.12526646, No 126011059). We thank anonymous reviewers for their insightful feedback.

References

[1] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *International Conference on Learning Representations - Workshops (ICLR-W)*, 2017.

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[4] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[6] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4619–4629, 2019.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[8] Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal alignment beyond cosine similarity. *arXiv preprint arXiv:2509.24734*, 2025.

[9] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[10] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[11] Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

[12] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[13] GroupLens research. MovieLens dataset, 2015.

[14] Yuhan Guo, cong guo, Aiwen Sun, Hongliang He, Xinyu Yang, Yue Lu, Yingji Zhang, Xuntao Guo, Dong Zhang, Jianzhuang Liu, Jiang Duan, Yijia Xiao, Liangjian Wen, Hai-Ming Xu, and Yong Dai. Webcogreasoner: Towards multimodal knowledge-induced cognitive reasoning for web agents. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 36186–36210, 2026.

[15] Zhaochen Guo, Zhixiang Shen, Xuanting Xie, Liangjian Wen, and Zhao Kang. Disentangling homophily and heterophily in multimodal graph clustering. *arXiv preprint arXiv:2507.15253*, 2025.

[16] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, 2019.

[17] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9995–10006, 2020.

[18] 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 *International Conference on Machine Learning (ICML)*, pages 4904–4916, 2021.

[19] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[20] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[21] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[22] 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*, 2025.

[23] Stefanos Koutoupis, Michaela Areti Zervou, Konstantinos Kontras, Maarten De Vos, Panagiotis Tsakalides, and Grigorios Tsagkatakis. The more, the merrier: Contrastive fusion for higher-order multimodal alignment. *arXiv preprint arXiv:2511.21331*, 2025.

[24] Pieter M Kroonenberg. *Applied multiway data analysis*. John Wiley & Sons, 2008.

[25] Michelle A Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3):582–596, 2020.

[26] Paul Pu Liang, Zihao Deng, Martin Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:32971–32998, 2023.

[27] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. MultiBench: Multiscale benchmarks for multimodal representation learning. In *Neural Information Processing Systems (NeurIPS) – Track on Datasets and Benchmarks*, volume 1, 2021.

[28] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, 2018.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[30] Osman Asif Malik and Stephen Becker. Low-rank tucker decomposition of large tensors using tensorsketch. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[31] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

[32] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision (ECCV)*, pages 529–544, 2022.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[34] Zoe Piran, Michal Klein, James Thornton, and Marco Cuturi. Contrasting multiple representations with the multi-marginal matching gap. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 40827–40842. PMLR, 21–27 Jul 2024.

[35] Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. *arXiv preprint arXiv:2503.11892*, 2025.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[37] Adriel Saporta, Aahlad Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. *Advances in Neural Information Processing Systems*, 37:56919–56957, 2024.

[38] Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In *Annual Conference on Learning Theory*, 2008.

[39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, pages 776–794, 2020.

[40] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems (NeurIPS)*, 33:6827–6839, 2020.

[41] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*, 2020.

[42] Charalampos E. Tsourakakis. MACH: fast randomized tensor decompositions. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2010, April 29 - May 1, 2010, Columbus, Ohio, USA*, pages 689–700. SIAM, 2010.

[43] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16041–16050, 2022.

[44] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pages 286–303. Springer, 2024.

[45] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[46] Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang, Jun Wang, Zenglin Xu, and Jiang Duan. Infmasking: Unleashing synergistic information by contrastive multimodal interactions. *arXiv preprint arXiv:2509.25270*, 2025.

[47] Liangjian Wen, Linjie Li, Jiang Duan, Yong Dai, Jianzhuang Liu, and Zhao Kang. Dependency, compression, and synergy: A unified information-theoretic view of multimodal learning. *arXiv preprint arXiv:2609.14421*, 2026.

[48] Liangjian Wen, Xiasi Wang, Jianzhuang Liu, and Zenglin Xu. Mveb: Self-supervised learning with multi-view entropy bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6097–6108, 2024.

[49] Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21501–21509, 2025.

[50] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

[51] 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*, 2026.

[52] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6995–7004, 2021.

[53] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1103–1114, 2017.

[54] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[55] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[56] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

Appendix

A Limitations and Future Work

Limitations. Our theoretical analysis relies on Assumption 1. In particular, the core results are stated under the pure-synergy condition in Definition 1, where the synergistic factor carries no mutual information with any proper subset of modalities once the shared factor is conditioned on. This assumption is useful for isolating the failure mode of pairwise contrastive learning and for deriving a clean connection between synergistic information and conditional total correlation. However, real multimodal data may exhibit mixed interaction structures: synergistic signals can coexist with residual pairwise dependence, partially redundant cues, or modality-specific shortcuts. In such cases, our theory can be interpreted as characterizing an idealized regime that clarifies when higher-order modeling is necessary. Although the empirical results in Sec. 4 and the InterSHAP analysis in Appendix D.1 suggest that the proposed regularization is beneficial beyond purely synthetic settings, a more systematic robustness study under approximate or partially overlapping synergy remains open.

Future Work. Future work can extend Definition 1 from pure synergy to approximate and mixed-synergy regimes. One promising direction is to develop data-dependent diagnostics for estimating the degree of redundancy, uniqueness, and synergy in real benchmarks, which could guide when higher-order regularization is most useful and how strongly it should be applied. Moreover, scaling the proposed objective to larger modality sets and integrating it with multimodal alignment objectives or large multimodal models [14] may further test whether synergy-aware learning improves general-purpose multimodal reasoning.

B Experimental Details

Implementation details. Unless otherwise specified, all methods are implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU with 24 GB of VRAM. We employ the AdamW optimizer [29] across all experiments. Each experimental configuration is trained over five independent runs using random seeds in the range $[42, 46]$. We report the mean performance along with the standard deviation. Models are trained for up to 100 epochs; the best checkpoint is selected based on validation accuracy, with early stopping applied when applicable to prevent overfitting. Detailed hyperparameter settings for each dataset and method are released alongside our source code.

Differentiability and Gradient Path of Tucker Decomposition. HRIL supports end-to-end backpropagation through the Tucker decomposition by leveraging the PyTorch autograd engine coupled with the TensorLy library's differentiable backend. We compute the approximate Tucker decomposition $\mathcal{T}_{r_p,v} \approx \mathcal{G}_v \times_1 A_{1,v} \cdots \times_M A_{M,v}$ using the Higher-Order Orthogonal Iteration (HOOI) algorithm [24], initialized via truncated SVD. The autodiff engine dynamically traces the alternating least-squares updates and embedded SVD operations executed during each forward pass. The solver employs a convergence tolerance (10^{-3}) on changes in reconstruction error alongside a maximum of 15 iterations, ensuring that the computational graph records the executed optimization steps. Consequently, gradients propagate backward through the actual HOOI trajectory of the executed finite iterations without truncation or manual intervention.

During backpropagation, the gradient with respect to the modality embeddings $\mathbf{Z}_{v,u}^{(m,b)}$ is computed via the chain rule as:

$$\frac{\partial \ell_{\text{syn},v}}{\partial Z_{v,u,a}^{(m,b)}} = \frac{1}{B} \sum_{j_1,\ldots,j_M} \mathcal{D}_{j_1 \ldots j_M} P_{a j_m}^{(m)} \prod_{k \neq m} Z_{p,v,j_k}^{(k,b)},$$

where the total derivative $\mathcal{D} = \partial \ell_{\text{syn},v} / \partial \mathcal{T}_{r_p,v}$ encompasses the full derivative chain through the traced HOOI steps and SVD differentials, $\ell_{\text{syn},v}$ denotes the per-view hinge in Eq. (5), and $\mathbf{Z}_{p,v}^{(m,b)} = \mathbf{Z}_{v,u}^{(m,b)} P^{(m)}$. This term is evaluated automatically by the framework via unrolled automatic differentiation, requiring no manual Jacobian derivation. Autograd also accounts for the factor $1/2$ from two-view averaging. The projection matrix $P^{(m)}$ is constructed once via QR decomposition of a random Gaussian matrix with a fixed seed and cached for both views and remains constant throughout training. As a static linear transformation, it introduces no additional optimization variables and provides a gradient path from the interaction core back to the trainable representation modules.

We provide the analysis of computational and optimization efficiency in Appendix D.6. Notably, the synergy-aware regularization $\mathcal{L}_{\text{syn}}$ employs a margin-based activation that, when numerical clamps are inactive, remains dormant in each view unless the core tensor energy concentration exceeds the prescribed threshold $1 - \xi$. The implementation floors $\|\mathcal{G}_v\|_F^2$ at 10^{-6}, clips the remaining-energy ratio to $[10^{-6}, 1 - 10^{-6}]$, and averages the two per-view hinges. This sparse activation mechanism sets the direct hinge gradient to zero when inactive, while the implementation still evaluates forward and backward operations through the Tucker decomposition module.

Model configuration. For dataset-specific encoder architectures, modality-specific data augmentations, and latent converters, we follow the same configurations as CoMM [11].

The fusion module in HRIL adopts a Transformer-based encoder to integrate modality-specific representations, following the design of CoMM [11]. Each layer comprises a multi-head self-attention mechanism followed by a position-wise feed-forward network, both equipped with residual connections and layer normalization. In bimodal settings, we employ a single-layer Transformer with 8 attention heads, whereas trimodal configurations utilize a 2-layer Transformer with the same head count. A learnable `[CLS]` token is prepended to the input sequence to aggregate cross-modal information into a unified global representation. For ConFu [23], we replicate the original implementation by using a 2-layer MLP as the subset fusion module.

C Dataset Details

C.1 Trifeature

The Trifeature dataset [17] serves as a controlled benchmark for probing feature learning in visual neural networks. It comprises three distinct attributes—shape, color, and texture—each instantiated with 10 categorical variants, yielding 1,000 unique compositions. We adopt the standard split of 800 compositions for training and 200 for evaluation. To enhance sample diversity, each training composition is augmented with three randomized instances, where shapes undergo independent rotations uniformly sampled from $[-45°, 45°]$. All shapes are rendered within a 128×128 region, then randomly positioned on a 224×224 canvas while preserving full visibility; color and texture are applied analogously. Following [11], we construct bimodal pairs from these instances, resulting in 10,000 training and 4,096 test pairs drawn *i.i.d.* from the same generative distribution.

C.2 Multibench

According to [27], all datasets below have been systematically de-identified to eliminate personally identifiable information and ensure strict privacy compliance. Multibench [27] provides access to the extracted features and labels for each dataset.

- **MIMIC** [20] is a large-scale critical care dataset comprising de-identified clinical records from over 38,000 patients admitted to intensive care units between 2001 and 2012. Each sample integrates two heterogeneous modalities: (i) a temporal modality consisting of 24-hour longitudinal clinical measurements (12-dimensional vectors), and (ii) a static tabular modality encoding demographic features (5-dimensional vectors).

  We follow the protocol of prior work [27, 26, 11] and formulate the task as binary disease classification: predicting whether a patient's diagnosis falls within ICD-9 (*International Statistical Classification of Diseases and Related Health Problems*) code group 7 (codes 460–519).
- **MOSI** [54] contains 2,199 YouTube-sourced video clips curated for multimodal sentiment analysis. Each sample comprises three modalities: visual, audio, and text. Ground-truth sentiment scores are originally annotated on a continuous scale from -3 to 3; following [26], these labels are binarized into positive and negative classes.

  In our experiments, we utilize the textual and visual modalities in bimodal configurations, while the trimodal setting incorporates all three modalities to evaluate the model's ability to integrate complementary information across modalities.
- **MOSEI** [55] is a large-scale multimodal benchmark for sentence-level sentiment analysis and emotion recognition, containing almost 23,000 monologue video clips. It comprises more than 65 hours of annotated video content, spanning 250 distinct categories and featuring

contributions from over 1,000 speakers. Each sample includes aligned visual, acoustic, and textual modalities, with sentiment labels originally annotated on a continuous scale from -3 to 3.

In our experiments, we use its sentiment annotations and, following [26], binarize them into positive and negative classes. We then evaluate the model's performance in trimodal settings.

- **UR-FUNNY** [16] is a large-scale benchmark for humor detection, curated from 1,866 TED talks and containing 16,514 multimodal samples. Each instance integrates visual frames, acoustic signals, and aligned textual transcripts, with the goal of discerning humorous versus non-humorous content.

  In our experiments, we utilize the textual and visual modalities in bimodal configurations, while the trimodal setting incorporates all three modalities.

- **MUStARD** [6] is a multimodal benchmark for sarcasm detection, compiled from clips of popular sitcoms such as *Friends*. The corpus contains 690 samples, each providing synchronized video, audio, and subtitle streams with binary sarcasm annotations.

  In our experiments, we utilize the textual and visual modalities in bimodal configurations.

- **Vision&Touch** [25] is a robotic manipulation benchmark with 150 trajectories of 1,000 time steps each. Its observations combine RGB and depth imagery with force measurements and end-effector position and velocity.

  In our experiments, we utilize the visual and proprioceptive modalities in a bimodal binary classification task to predict whether contact will occur at the next time step.

C.3 MM-IMDb

Multimodal IMDb (MM-IMDb) [2] poses movie-genre prediction as a multilabel classification task with 23 categories. Its 25,959 movies are drawn from MovieLens 20M [13] and include posters, plot descriptions, genre annotations, and metadata. We use the posters and plot descriptions as the image and text modalities, respectively, and evaluate weighted and macro F1. Although the dataset is included in MultiBench [27], our experiments train on its original image and text inputs.

D Additional Experiments

D.1 Evaluations with InterSHAP

Metric definition. Rather than scoring only downstream task performance, this measure inspects how predictions are formed by separating cross-modal interaction effects from modality-specific main effects. Concretely, predictions $f(x)$ are decomposed with the Shapley Interaction Index (SII), where pairwise terms represent interaction strength and diagonal terms represent unimodal contributions. Let M be the number of modalities and let ϕ_{ij} denote the SII attribution between modalities i and j (computed either per sample or aggregated over the dataset). The InterSHAP score [49] is then given by:

$$\text{InterSHAP} = \frac{\sum_{\substack{i,j=1 \\ i \neq j}}^{M} |\phi_{ij}|}{\sum_{i,j=1}^{M} |\phi_{ij}|}, \tag{9}$$

where the numerator aggregates off-diagonal interaction attributions ($i \neq j$) and the denominator aggregates all absolute attributions, including diagonal main effects ($i = j$). Larger values indicate that model decisions rely more on interaction terms than on unimodal effects.

Table 7 extends the comparison in Table 5 with CoMM and a multilinear-rank control on the UR-FUNNY and MOSI datasets, where the multilinear-rank control replaces $\mathcal{L}_{\text{syn}}$ with a low-rank penalty on the Tucker core while keeping the other losses unchanged. We use this control to test whether a multilinear-rank restriction alone explains the mean InterSHAP gains. Its loss is $\mathcal{L}_{\text{rank}} = \sum_{m=1}^{M} \sum_{i>k_m} \sigma_i^2(\mathcal{G}_{(m)})$. The penalty sums the squared singular values beyond the leading k_m singular values of each mode unfolding $\mathcal{G}_{(m)}$. This control encourages low multilinear rank, whereas $\mathcal{L}_{\text{syn}}$ activates only when one core coordinate exceeds the prescribed energy share.

Table 7: Further comparison with InterSHAP on the UR-FUNNY and MOSI datasets.

Model	*UR-FUNNY*	*MOSI*
CoMM	$7.77_{\pm 3.05}$	$3.82_{\pm 1.90}$
HRIL without $\mathcal{L}_{\text{syn}}$	$7.78_{\pm 4.56}$	$6.82_{\pm 3.73}$
Multilinear-rank control	$8.17_{\pm 6.33}$	$5.87_{\pm 4.26}$
HRIL	$9.25_{\pm 5.56}$	$9.90_{\pm 4.81}$

As shown in Table 7, HRIL has the highest mean InterSHAP score on both datasets. The multilinear-rank control penalty yields a small mean increase over the variant without synergy regularization on UR-FUNNY but a decrease on MOSI. HRIL exceeds it by 1.08 and 4.03 percentage points, respectively. These results support preserving higher-order interaction capacity within the prescribed Tucker subspace. The regularizer $\mathcal{L}_{\text{syn}}$ discourages excessive dominance by a single core coordinate while allowing the core to remain sparse.

D.2 Hyperparameter Sensitivity

Table 8 reports sensitivity analyses on trimodal UR-FUNNY and MOSEI, with the remaining hyperparameters fixed. We vary the Tucker rank r_m, the projection dimension r_p and the margin ξ. The Tucker rank determines the retained dimension along each mode, while r_p determines the number of entries r_p^M in the cross-moment tensor. The margin ξ sets the threshold $1 - \xi$ above which core-energy concentration is penalized.

Table 8: Sensitivity analysis: accuracy (%)

Hyperparameter	*Value*	*UR-FUNNY*	*MOSEI*
Tucker rank r_m	8	$65.0_{\pm 0.54}$	$69.9_{\pm 0.31}$
	16	$64.9_{\pm 1.34}$	$70.0_{\pm 0.47}$
	24	$65.5_{\pm 0.56}$	$70.5_{\pm 0.20}$
Projection dimension r_p	8	$64.8_{\pm 0.59}$	$70.2_{\pm 0.26}$
	16	$64.9_{\pm 0.66}$	$70.2_{\pm 0.20}$
	24	$65.5_{\pm 0.56}$	$70.5_{\pm 0.20}$
	32	$65.5_{\pm 0.90}$	$70.2_{\pm 0.45}$
Margin ξ	0.1	$65.2_{\pm 1.15}$	$70.0_{\pm 0.40}$
	0.3	$65.2_{\pm 1.05}$	$70.5_{\pm 0.38}$
	0.5	$65.5_{\pm 0.56}$	$70.5_{\pm 0.20}$
	0.7	$65.2_{\pm 1.00}$	$70.3_{\pm 0.25}$

As shown in Table 8, a Tucker rank of 24 gives the highest mean accuracy on both datasets. A projection dimension of 24 also attains the highest mean accuracy in its separate sweep, while increasing it to 32 enlarges the tensor without further mean improvement. Since the tensor is constructed before Tucker decomposition, reducing the decomposition rank does not lower the tensor construction cost. The margin sweep is comparatively stable: mean accuracy varies by 0.3 percentage points on UR-FUNNY and 0.5 on MOSEI, with $\xi = 0.5$ among the best settings on both datasets.

D.3 Batch Size Robustness

Since the synergy-aware regularization is computed from the empirical cross-moment tensor (Eq. (4)), we further evaluate the sensitivity of HRIL to the mini-batch size. As shown in Fig. 3, HRIL demonstrates more pronounced robustness on datasets with larger sample sizes. This trend is expected because larger datasets provide more reliable mini-batch estimates of cross-modal moments, whereas smaller datasets may introduce higher estimation variance. In addition, HRIL achieves its best performance with a batch size of 64, and further increasing the batch size does not lead to consistent gains. These results indicate that HRIL does not require excessively large batches for more reliable estimates of cross-modal moments, while maintaining both scalability and computational efficiency.

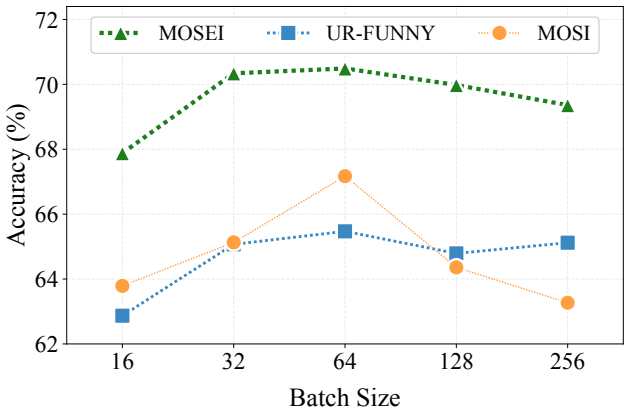

Figure 3: Performance sensitivity to mini-batch size on MOSEI, UR-FUNNY, and MOSI.

D.4 Experiments on MM-IMDb

We further evaluate HRIL on MM-IMDb multilabel movie-genre prediction using the original posters and plot descriptions. Table 9 reports linear-probing results, where V and L denote visual and textual inputs. The CLIP results first show that combining posters and plot descriptions improves over either modality alone, indicating that the joint input provides additional predictive value. With the same backbone, HRIL achieves weighted and macro F1 scores of 62.30% and 55.44%, exceeding CoMM by 1.01 and 1.65 percentage points, respectively. Thus, by considering multilabel classification directly from raw image and text inputs, this experiment broadens the evaluation beyond the original MultiBench experiments and further demonstrates the effectiveness of HRIL.

Table 9: Linear probing F1 scores (%) on MM-IMDb. $^{\triangle}$ indicates further training on unlabeled data. $^{\clubsuit}$ denotes results from [11].

Model	*Modalities*	*weighted-F1*↑	*macro-F1*↑
SimCLR$^{\clubsuit\triangle}$ [7]	V	$40.35_{\pm 0.23}$	$27.99_{\pm 0.33}$
CLIP$^{\clubsuit}$ [36]	V	51.5	40.8
	L	51.0	43.0
	V+L	58.9	50.9
SLIP$^{\clubsuit\triangle}$ [32]	V+L	$56.54_{\pm 0.19}$	$47.35_{\pm 0.27}$
CLIP$^{\clubsuit\triangle}$ [36]	V+L	$54.49_{\pm 0.19}$	$44.94_{\pm 0.30}$
CoMM$_{\text{(CLIP backbone)}}$	V+L	$61.29_{\pm 0.73}$	$53.79_{\pm 0.22}$
HRIL$_{\text{(ours, CLIP backbone)}}$	V+L	$\mathbf{62.30}_{\pm 0.26}$	$\mathbf{55.44}_{\pm 0.26}$

D.5 Experiments on Vision&Touch

To evaluate the effectiveness of HRIL on a dataset with a larger sample size, we further conduct experiments on Vision&Touch [25], which contains 147,000 samples. Given visual and proprioceptive observations, the task is to predict whether contact will occur at the next time step.

Table 10: Linear probing accuracy (%) on Vision&Touch contact prediction (V&T CP).

Model	*V&T CP*↑
Cross	$86.3_{\pm 0.25}$
Cross+Self	$87.6_{\pm 0.26}$
CoMM	$87.0_{\pm 1.77}$
HRIL (ours)	$\mathbf{88.6}_{\pm 0.14}$

As shown in Table 10, adding the self-contrastive objective to Cross raises contact-prediction accuracy from 86.3% to 87.6%, making Cross+Self the strongest listed baseline. Our method reaches 88.6%, improving on Cross+Self by 1.0 percentage point and on CoMM by 1.6 points. These results further demonstrate the effectiveness of HRIL on a substantially larger dataset for robotic contact prediction.

D.6 Computational Complexity and Runtime Analysis

Table 11: Training efficiency comparison between CoMM [11] and HRIL on trimodal Multibench [27] datasets. Both methods use the same multimodal backbone. We report the average wall-clock step time, optimizer update time, and the percentage of each step spent in backward computation.

Dataset	*Model*	*Training Efficiency Metrics*		
		Step Time (ms)	*Optimizer Time (ms)*	*Backward / Step Ratio (%)*
UR-FUNNY	CoMM	90.73	3.86	53.82
	HRIL	155.65	3.47	46.23
MOSI	CoMM	90.00	3.90	54.07
	HRIL	156.95	3.47	46.96
MOSEI	CoMM	90.57	4.06	53.99
	HRIL	159.25	3.63	47.56

As shown in Tab. 11, HRIL increases the per-step runtime from roughly 90 ms to 156–159 ms across the three datasets compared with CoMM [11]. The optimizer time remains nearly unchanged, indicating that the overhead does not come from additional trainable parameters or a heavier parameter update. Instead, the extra cost mainly comes from the forward and backward computation of the additional loss terms, especially the cross-moment tensor regularization in Eq. (5). For a mini-batch of B samples and M modalities, each projected to r_p dimensions, explicit construction of the cross-moment tensor takes $O(Br_p^M)$ time, and the resulting tensor occupies $O(r_p^M)$ storage before decomposition. With the batch size and model configuration fixed, increasing the dataset size requires more batches per epoch without changing the loss computation per batch.

These additional computations are confined to training, as the synergy-aware regularizer is not used during inference. Given the performance improvements in the evaluated settings, we regard the additional training cost as a reasonable trade-off. Potential ways to reduce this cost include TensorSketch-based Tucker approximation [30], which uses compressed representations in Tucker computations, and randomized tensor sparsification [42], which samples and reweights tensor entries before decomposition. Applying these approximations to our Tucker-core regularizer and evaluating their effects on training remain future work.

E Broader Impact

This work advances multimodal representation learning by explicitly preserving synergistic interactions across heterogeneous data sources. Its primary societal benefit lies in enabling more robust fusion frameworks for high-stakes domains such as clinical decision support, where joint modality reasoning is critical. Given the methodological nature of the contribution, direct societal risks are minimal; however, we emphasize that downstream deployment should adhere to established safeguards regarding data privacy, algorithmic fairness, and computational sustainability.

F Pseudo-code

Algorithm 1 summarizes the training procedure of HRIL, using the same notation as the main paper. Here $D = \sum_{m=1}^{M} d_m$. We use the InfoNCE mutual-information estimator [33] defined in the main paper:

$$\hat{I}_{\text{NCE}}(\mathbf{Z}, \mathbf{Z}') = \mathbb{E}_{\substack{\mathbf{z},\mathbf{z}'_{\text{pos}} \sim p(\mathbf{Z},\mathbf{Z}') \\ \mathbf{z}'_{\text{neg}} \sim p(\mathbf{Z}')}} \left[\log \frac{\exp \text{sim}(\mathbf{z}, \mathbf{z}'_{\text{pos}})}{\exp \text{sim}(\mathbf{z}, \mathbf{z}'_{\text{pos}}) + \sum_{\mathbf{z}'_{\text{neg}}} \exp \text{sim}(\mathbf{z}, \mathbf{z}'_{\text{neg}})} \right],$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

Algorithm 1 HRIL training algorithm

Require: Multimodal dataset $\{X_1, \dots, X_M\}$, label-preserving transformations $\mathcal{T}^\star$, mini-batch size B, unimodal encoders $\{f_m\}_{m=1}^M$, shared Transformer g, QR projection matrices $\{P^{(m)}\}_{m=1}^M$, projection dimension r_p, Tucker ranks $(r_1, \dots, r_M)$, margin ξ, alignment hyperparameters γ, ϵ, loss weights λ_a, λ_s.

for each sampled mini-batch $\{(\mathbf{x}^{(1,b)}, \dots, \mathbf{x}^{(M,b)})\}_{b=1}^B$ **do**
  draw $t_1, t_2 \sim \mathcal{T}^\star$
  for $b \leftarrow 1$ to B **do**
    for $m \leftarrow 1$ to M **do**
      $\mathbf{x}_1^{(m,b)}, \mathbf{x}_2^{(m,b)} \leftarrow t_1(\mathbf{x}^{(m,b)}), t_2(\mathbf{x}^{(m,b)})$
      $\mathbf{h}_1^{(m,b)}, \mathbf{h}_2^{(m,b)} \leftarrow f_m(\mathbf{x}_1^{(m,b)}), f_m(\mathbf{x}_2^{(m,b)})$
      $\mathbf{Z}_{1,u}^{(m,b)}, \mathbf{Z}_{2,u}^{(m,b)} \leftarrow g(\mathbf{h}_1^{(m,b)}), g(\mathbf{h}_2^{(m,b)})$
      $\mathbf{Z}_{p,1}^{(m,b)}, \mathbf{Z}_{p,2}^{(m,b)} \leftarrow \mathbf{Z}_{1,u}^{(m,b)} P^{(m)}, \mathbf{Z}_{2,u}^{(m,b)} P^{(m)}$
    end for
    $\mathbf{Z}_{1,f}^{(b)} \leftarrow g(\mathbf{h}_1^{(1,b)}, \dots, \mathbf{h}_1^{(M,b)})$
    $\mathbf{Z}_{2,f}^{(b)} \leftarrow g(\mathbf{h}_2^{(1,b)}, \dots, \mathbf{h}_2^{(M,b)})$
  end for
  for $m \leftarrow 1$ to M **do**
    $\mathcal{L}_{\text{InfoNCE}}^{(m)} \leftarrow -\frac{1}{2}\Big(\hat{I}_{\text{NCE}}(\mathbf{Z}_{1,u}^{(m)}, \mathbf{Z}_{2,f}) + \hat{I}_{\text{NCE}}(\mathbf{Z}_{2,u}^{(m)}, \mathbf{Z}_{1,f})\Big)$
  end for
  for each view $v \in \{1, 2\}$ **do**
    $\mathcal{T}_{r_p,v} \leftarrow \frac{1}{B}\sum_{b=1}^B \mathbf{Z}_{p,v}^{(1,b)} \otimes \cdots \otimes \mathbf{Z}_{p,v}^{(M,b)}$
    Tucker decompose $\mathcal{T}_{r_p,v} \approx \mathcal{G}_v \times_1 \mathbf{A}_{1,v} \cdots \times_M \mathbf{A}_{M,v}$ with ranks $(r_1, \dots, r_M)$
    $\rho(\mathcal{G}_v) \leftarrow \max_{i_1,\dots,i_M} \mathcal{G}_{i_1 \dots i_M,v}^2 / \|\mathcal{G}_v\|_F^2$
    $\tilde{\mathbf{Z}}_v^{(b)} \leftarrow [\tilde{\mathbf{Z}}_v^{(1,b)}; \dots; \tilde{\mathbf{Z}}_v^{(M,b)}]$ for all b, where $\tilde{\mathbf{Z}}_v^{(m,b)} := \mathbf{Z}_{v,u}^{(m,b)} / \|\mathbf{Z}_{v,u}^{(m,b)}\|$
    $\hat{\mathbf{Z}}_v^{(b)} \leftarrow \tilde{\mathbf{Z}}_v^{(b)} / \|\tilde{\mathbf{Z}}_v^{(b)}\|$ for all b
  end for
  $\mathcal{L}_{\text{syn}} \leftarrow \Big([\xi - (1 - \rho(\mathcal{G}_1))]^+ + [\xi - (1 - \rho(\mathcal{G}_2))]^+\Big)/2$
  $\mathcal{L}_{\text{inv}} \leftarrow \frac{1}{B}\sum_{b=1}^B \big(1 - \text{sim}(\hat{\mathbf{Z}}_1^{(b)}, \hat{\mathbf{Z}}_2^{(b)})\big)^2$
  $\bar{\mathbf{Z}} \leftarrow (\hat{\mathbf{Z}}_1 + \hat{\mathbf{Z}}_2)/2$
  $\mathcal{L}_{\text{var}} \leftarrow \frac{1}{D}\sum_{d=1}^D \Big[\gamma - \sqrt{\text{Var}(\bar{\mathbf{Z}}_d) + \epsilon}\Big]^+$
  $\mathcal{L}_{\text{align}} \leftarrow \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{var}}$
  $\mathcal{L}_{\text{total}} \leftarrow \sum_{m=1}^M \mathcal{L}_{\text{InfoNCE}}^{(m)} + \lambda_a \mathcal{L}_{\text{align}} + \lambda_s \mathcal{L}_{\text{syn}}$
  update $(f_m)_{m\in[1..M]}$ and g to minimize $\mathcal{L}_{\text{total}}$
end for
return $(f_m)_{m\in[1..M]}, g$

G Proofs

Theorem 1 (Insufficiency of Pairwise Minimal Sufficient Representation)**.** *Under Assumption 1, suppose that* $\mathbf{q}_{syn}$ *is a pure synergistic factor in the sense of Definition 1 and that* $I(Y; \mathbf{q}_{syn} \mid \mathbf{q}_r, \mathbf{q}_{1:M}) > 0$*. Then*

$$I(Y; \mathbf{Z}_{pair}^{\min}) \le I(Y; \mathbf{q}_r).$$

Moreover, the total information gap is lower bounded as

$$I(Y; X_{1:M}) - I(Y; \mathbf{Z}_{pair}^{\min}) \;\ge\; \underbrace{I(Y; \mathbf{q}_{1:M} \mid \mathbf{q}_r)}_{\textit{unique gap}} + \underbrace{I(Y; \mathbf{q}_{syn} \mid \mathbf{q}_r, \mathbf{q}_{1:M})}_{\textit{synergistic gap}}.$$

Proof. We prove the two claims in order.

Step 1: pairwise alignment captures at most $\mathbf{q}_r$.

(a) Pure synergy is inaccessible to individual modality pairs. For any pair $i \neq j$ with $M \geq 3$, the set $\{i, j\}$ is a proper subset of the modalities. Definition 1 therefore gives

$$I(\mathbf{q}_{\text{syn}}; X_i, X_j \mid \mathbf{q}_r) = 0.$$

Thus, a modality pair contains no information about the pure synergistic factor once the shared factor is given.

(b) The pairwise-minimal representation excludes non-shared information. For $\mathbf{Z}_{\text{pair}}^{\min}$, we use the shared information characterization in Eq. (2). The encoding relation $\mathbf{Z} = f(X)$ in the task setup ensures that a representation is a function of the multimodal observation. Conditional data processing consequently gives

$$\begin{aligned} 0 &\leq I(\mathbf{Z}_{\text{pair}}^{\min}; \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}} \mid \mathbf{q}_r) \\ &\leq I(\mathbf{Z}_{\text{pair}}^{\min}; X_{1:M} \mid \mathbf{q}_r) = 0, \end{aligned}$$

where the last equality is Eq. (2). Hence,

$$\mathbf{Z}_{\text{pair}}^{\min} \perp (\mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}}) \mid \mathbf{q}_r.$$

(c) Upper bound on task information. Assumption 1 states that $Y \perp X_{1:M} \mid (\mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}})$. Since $\mathbf{Z}_{\text{pair}}^{\min}$ is a function of $X_{1:M}$, this also implies

$$Y \perp \mathbf{Z}_{\text{pair}}^{\min} \mid (\mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}}).$$

Combining this conditional independence with part (b), the chain rule gives

$$\begin{aligned} &I(\mathbf{Z}_{\text{pair}}^{\min}; Y, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}} \mid \mathbf{q}_r) \\ &\quad = I(\mathbf{Z}_{\text{pair}}^{\min}; \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}} \mid \mathbf{q}_r) + I(\mathbf{Z}_{\text{pair}}^{\min}; Y \mid \mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}}) \\ &\quad = 0. \end{aligned}$$

By nonnegativity of conditional mutual information, $I(Y; \mathbf{Z}_{\text{pair}}^{\min} \mid \mathbf{q}_r) = 0$. Another application of the chain rule yields

$$\begin{aligned} I(Y; \mathbf{Z}_{\text{pair}}^{\min}) &= I(Y; \mathbf{q}_r) + I(Y; \mathbf{Z}_{\text{pair}}^{\min} \mid \mathbf{q}_r) - I(Y; \mathbf{q}_r \mid \mathbf{Z}_{\text{pair}}^{\min}) \\ &\leq I(Y; \mathbf{q}_r). \end{aligned}$$

Step 2: three-layer gap decomposition. By the chain rule applied to the latent factors,

$$\begin{aligned} I(Y; \mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}}) = I(Y; \mathbf{q}_r) &+ I(Y; \mathbf{q}_{1:M} \mid \mathbf{q}_r) \\ &+ I(Y; \mathbf{q}_{\text{syn}} \mid \mathbf{q}_r, \mathbf{q}_{1:M}). \end{aligned}$$

Assumption 1 and the condition stated immediately after it give, respectively,

$$\begin{aligned} I(Y; X_{1:M} \mid \mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}}) &= 0, \\ I(Y; \mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}} \mid X_{1:M}) &= 0. \end{aligned}$$

Expanding $I(Y; X_{1:M}, \mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}})$ in the two possible orders therefore yields

$$I(Y; X_{1:M}) = I(Y; \mathbf{q}_r, \mathbf{q}_{1:M}, \mathbf{q}_{\text{syn}}).$$

Together with the bound in Step 1, this gives

$$\begin{aligned} I(Y; X_{1:M}) - I(Y; \mathbf{Z}_{\text{pair}}^{\min}) \geq I(Y; \mathbf{q}_{1:M} \mid \mathbf{q}_r) &\\ + I(Y; \mathbf{q}_{\text{syn}} \mid \mathbf{q}_r, \mathbf{q}_{1:M}). & \end{aligned}$$

The first term is the unique gap; the second is the synergistic gap. The second term is strictly positive by hypothesis, so the total information gap is strictly positive. □

Proposition 1 (Conditional Total Correlation Decomposition)**.** *For M modalities, the conditional total correlation admits the chain-rule decomposition*

$$\text{TC}(X_{1:M} \mid \mathbf{q}_r) = \sum_{k=2}^{M} I(X_k;\, X_1, \ldots, X_{k-1} \mid \mathbf{q}_r).$$

In particular, for $M = 3$: $\text{TC}(X_1, X_2, X_3 \mid \mathbf{q}_r) = I(X_1; X_2 \mid \mathbf{q}_r) + I(X_3; X_1, X_2 \mid \mathbf{q}_r)$, and symmetrically for all permutations.

Proof. We provide the proof for $M = 3$, as the extension to general M is straightforward.

By definition,

$$\mathrm{TC}(X_1, X_2, X_3 \mid \mathbf{q}_r) = H(X_1 \mid \mathbf{q}_r) + H(X_2 \mid \mathbf{q}_r) + H(X_3 \mid \mathbf{q}_r) - H(X_1, X_2, X_3 \mid \mathbf{q}_r).$$

Using the chain rule,

$$H(X_1, X_2, X_3 \mid \mathbf{q}_r) = H(X_1 \mid \mathbf{q}_r) + H(X_2 \mid X_1, \mathbf{q}_r) + H(X_3 \mid X_1, X_2, \mathbf{q}_r).$$

Substituting this identity yields

$$\mathrm{TC}(X_1, X_2, X_3 \mid \mathbf{q}_r) = I(X_1; X_2 \mid \mathbf{q}_r) + I(X_3; X_1, X_2 \mid \mathbf{q}_r).$$

The symmetric form follows by applying the chain rule under each permutation. □

Theorem 2 (Synergy Total Correlation Identity)**.** *Under Definition 1, the synergistic information satisfies the identity*

$$I(\mathbf{q}_{syn}; X_{1:M} \mid \mathbf{q}_r) = \mathrm{TC}(X_{1:M} \mid \mathbf{q}_r, \mathbf{q}_{syn}) - \mathrm{TC}(X_{1:M} \mid \mathbf{q}_r).$$

Proof. We first establish the above equation and further interpret this theorem.

Step 1: the identity. By the definition of mutual information,

$$I(\mathbf{q}_{\text{syn}}; X_{1:M} \mid \mathbf{q}_r) = H(X_{1:M} \mid \mathbf{q}_r) - H(X_{1:M} \mid \mathbf{q}_r, \mathbf{q}_{\text{syn}}).$$

Recall that $\mathrm{TC}(X_{1:M} \mid C) := \sum_m H(X_m \mid C) - H(X_{1:M} \mid C)$ for any conditioning variable C. Rewriting both joint entropies:

$$H(X_{1:M} \mid \mathbf{q}_r) = \textstyle\sum_m H(X_m \mid \mathbf{q}_r) - \mathrm{TC}(X_{1:M} \mid \mathbf{q}_r),$$
$$H(X_{1:M} \mid \mathbf{q}_r, \mathbf{q}_{\text{syn}}) = \textstyle\sum_m H(X_m \mid \mathbf{q}_r, \mathbf{q}_{\text{syn}}) - \mathrm{TC}(X_{1:M} \mid \mathbf{q}_r, \mathbf{q}_{\text{syn}}).$$

Definition 1 gives $I(\mathbf{q}_{\text{syn}}; X_m \mid \mathbf{q}_r) = 0$ for every singleton $\{m\} \subsetneq \{1, \ldots, M\}$, so $H(X_m \mid \mathbf{q}_r, \mathbf{q}_{\text{syn}}) = H(X_m \mid \mathbf{q}_r)$. Substituting and cancelling:

$$I(\mathbf{q}_{\text{syn}}; X_{1:M} \mid \mathbf{q}_r) = \mathrm{TC}(X_{1:M} \mid \mathbf{q}_r, \mathbf{q}_{\text{syn}}) - \mathrm{TC}(X_{1:M} \mid \mathbf{q}_r).$$

Step 2: proper-subset marginals remain unchanged. Definition 1 gives $\mathbf{q}_{\text{syn}} \perp X_A \mid \mathbf{q}_r$ for every proper subset $A \subsetneq \{1, \ldots, M\}$. Hence conditioning on $\mathbf{q}_{\text{syn}}$ leaves all proper-subset marginals unchanged, and the TC difference comes from the change in the full joint distribution. For $M \geq 3$, this includes every pairwise marginal. □

Proposition 2 (Shared-Factor Dominance Implies Redundant Low-Order Structure)**.** *Under Assumption 1 and the representation model (Eq.* (1)*), let* $W_{ij}^* = \arg\min_W \mathbb{E}\|\mathbf{Z}_i - W\mathbf{Z}_j\|^2$ *be the population-optimal cross-modal linear predictor. Then:*

- ***(a) (Rank bound)*** $\mathrm{rank}(W_{ij}^*) \leq d_r$*, where* $d_r = \dim(\mathbf{c}_r)$ *is the dimension of the shared subspace.*
- ***(b) (Collapse under dominance)*** *When the shared-to-total energy ratio approaches one, the normalized reconstruction error vanishes and the embeddings exhibit a redundant low-order structure.*

Proof. **(a) Rank bound.** Under the model $\mathbf{Z}_m = V_m^r \mathbf{c}_r + V_m^u \mathbf{c}_m + \boldsymbol{\epsilon}_m$ with mutually uncorrelated components, the cross-covariance is

$$\mathrm{Cov}(\mathbf{Z}_i, \mathbf{Z}_j) = V_i^r \Sigma_r (V_j^r)^\top,$$

where $\Sigma_r := \mathrm{Cov}(\mathbf{c}_r) \succ 0$. This matrix has rank at most $\min(\mathrm{rank}(V_i^r), \mathrm{rank}(V_j^r)) \leq d_r$. The auto-covariance of $\mathbf{Z}_j$ is

$$\Sigma_j := \mathrm{Cov}(\mathbf{Z}_j) = V_j^r \Sigma_r (V_j^r)^\top + V_j^u \Sigma_j^u (V_j^u)^\top + \Sigma_{\boldsymbol{\epsilon}_j} \succ 0,$$

where $\Sigma_j^u := \mathrm{Cov}(\mathbf{c}_j)$ and $\Sigma_{\boldsymbol{\epsilon}_j} := \mathrm{Cov}(\boldsymbol{\epsilon}_j)$. Positivity follows from the noise term. The population-optimal predictor is

$$W_{ij}^* = V_i^r \Sigma_r (V_j^r)^\top \Sigma_j^{-1},$$

whose rank is at most d_r since $\mathrm{col}(W_{ij}^*) \subseteq \mathrm{col}(V_i^r)$.

(b) Collapse under dominance. For simplicity, we assume that the observed vectors have been centered and normalized sample-wise: $\mathbb{E}[\mathbf{Z}_m] = \mathbf{0}$ and $\mathbb{E}\|\mathbf{Z}_m\|^2 = \mathrm{tr}(\Sigma_m) = 1$. Define the unique-plus-noise component of modality m as

$$\mathbf{n}_m \coloneqq V_m^u \mathbf{c}_m + \boldsymbol{\epsilon}_m, \qquad \Sigma_{\mathbf{n}_m} \coloneqq \mathrm{Cov}(\mathbf{n}_m) = V_m^u \Sigma_m^u (V_m^u)^\top + \Sigma_{\boldsymbol{\epsilon}_m} \succeq 0.$$

Then $\mathbf{Z}_m = V_m^r \mathbf{c}_r + \mathbf{n}_m$ with $\mathrm{Cov}(\mathbf{c}_r, \mathbf{n}_m) = \mathbf{0}$. We impose the following regularity conditions:

(A1) $\Sigma_{\mathbf{n}_j} \succ 0$ (or at least $\Sigma_j \succ 0$) to ensure the existence of the optimal linear predictor.

(A2) The unique-plus-noise components of different modalities are mutually uncorrelated: $\mathrm{Cov}(\mathbf{n}_i, \mathbf{n}_j) = \mathbf{0}$ for $i \neq j$.

(A3) V_j^r has full column rank d_r and $\Sigma_r \succ 0$.

Because $\mathrm{tr}(\Sigma_m) = 1$, the *shared energy ratio* simplifies to

$$\alpha_m \coloneqq \mathrm{tr}\big(V_m^r \Sigma_r (V_m^r)^\top\big) \in [0,1], \qquad \text{and} \qquad \mathrm{tr}(\Sigma_{\mathbf{n}_m}) = 1 - \alpha_m.$$

Hence $\alpha_m \to 1$ implies $\mathrm{tr}(\Sigma_{\mathbf{n}_m}) \to 0$. Since $\Sigma_{\mathbf{n}_m} \succeq 0$,

$$\|\Sigma_{\mathbf{n}_m}\|_2 \leq \mathrm{tr}(\Sigma_{\mathbf{n}_m}) \to 0.$$

The optimal linear predictor of $\mathbf{c}_r$ given $\mathbf{Z}_j$ is

$$\widehat{\mathbf{c}}_r = \Sigma_r (V_j^r)^\top \Sigma_j^{-1} \mathbf{Z}_j, \qquad \Omega_j \coloneqq \mathrm{Cov}(\mathbf{c}_r - \widehat{\mathbf{c}}_r) = \Sigma_r - \Sigma_r (V_j^r)^\top \Sigma_j^{-1} V_j^r \Sigma_r,$$

where $\Sigma_j = V_j^r \Sigma_r (V_j^r)^\top + \Sigma_{\mathbf{n}_j} \succ 0$.

To prove $\Omega_j \to \mathbf{0}$ rigorously, we bound it by a feasible sub-optimal estimator. By (A3), let $L_j = \big((V_j^r)^\top V_j^r\big)^{-1} (V_j^r)^\top$ be a left inverse of V_j^r. Define $\check{\mathbf{c}}_r = L_j \mathbf{Z}_j$. Then

$$\mathbf{c}_r - \check{\mathbf{c}}_r = -L_j \mathbf{n}_j, \qquad \mathrm{Cov}(\mathbf{c}_r - \check{\mathbf{c}}_r) = L_j \Sigma_{\mathbf{n}_j} L_j^\top.$$

By the optimality of the MMSE predictor, $\Omega_j \preceq \mathrm{Cov}(\mathbf{c}_r - \check{\mathbf{c}}_r)$. Taking spectral norms,

$$\|\Omega_j\|_2 \leq \|L_j\|_2^2 \, \|\Sigma_{\mathbf{n}_j}\|_2 \to 0,$$

so $\Omega_j \to \mathbf{0}$.

The best linear predictor of $\mathbf{Z}_i$ from $\mathbf{Z}_j$ is

$$W_{ij}^* = \mathrm{Cov}(\mathbf{Z}_i, \mathbf{Z}_j)\, \Sigma_j^{-1} = V_i^r \Sigma_r (V_j^r)^\top \Sigma_j^{-1},$$

whose column space lies in $\mathrm{col}(V_i^r)$, hence $\mathrm{rank}(W_{ij}^*) \leq d_r$. Its normalized mean-squared error can be written as

$$\frac{\mathbb{E}\|\mathbf{Z}_i - W_{ij}^* \mathbf{Z}_j\|^2}{\mathbb{E}\|\mathbf{Z}_i\|^2} = 1 - \frac{\mathrm{tr}\big(V_i^r \Sigma_r (V_j^r)^\top \Sigma_j^{-1} V_j^r \Sigma_r (V_i^r)^\top\big)}{\mathrm{tr}(\Sigma_i)}.$$

Using $\Sigma_r (V_j^r)^\top \Sigma_j^{-1} V_j^r \Sigma_r = \Sigma_r - \Omega_j$,

$$\mathrm{tr}\big(V_i^r \Sigma_r (V_j^r)^\top \Sigma_j^{-1} V_j^r \Sigma_r (V_i^r)^\top\big) = \mathrm{tr}\big(V_i^r \Sigma_r (V_i^r)^\top\big) - \mathrm{tr}\big(V_i^r \Omega_j (V_i^r)^\top\big) = \alpha_i - \mathrm{tr}\big(V_i^r \Omega_j (V_i^r)^\top\big).$$

The second term satisfies

$$0 \leq \mathrm{tr}\big(V_i^r \Omega_j (V_i^r)^\top\big) \leq \|V_i^r\|_2^2 \, \|\Omega_j\|_2 \, \mathrm{rank}(V_i^r) \to 0.$$

Because the normalization guarantees $\mathrm{tr}(\Sigma_i) = 1$, we obtain

$$\frac{\mathbb{E}\|\mathbf{Z}_i - W_{ij}^* \mathbf{Z}_j\|^2}{\mathbb{E}\|\mathbf{Z}_i\|^2} = 1 - \alpha_i + o(1).$$

If $\alpha_i \to 1$ and $\alpha_j \to 1$ (the latter ensures $\Omega_j \to \mathbf{0}$), the right-hand side tends to 0.

□

Lemma 1 (Redundancy Implies Low Multilinear Complexity)**.** *If the embeddings exhibit a redundant low-order structure in the sense of Definition 2, then the cross-moment tensor $\mathcal{T} = \mathbb{E}[\bigotimes_m \mathbf{Z}_m]$ admits a low multilinear-complexity approximation. In particular, the Tucker core $\mathcal{G}$ of $\mathcal{T}$ has at least one mode whose effective multilinear rank satisfies:*

$$\operatorname{rank}_{(k)}(\mathcal{G}) \leq \sum_{j \neq i} r_{ij} \quad \textit{for modality } k = i.$$

Proof. Without loss of generality, assume modality 1 satisfies the redundancy condition. By Definition 2, there exist linear maps $\mathbf{W}_{1j}$ such that

$$\mathbf{Z}_1 = \sum_{j=2}^{M} \mathbf{W}_{1j}\mathbf{Z}_j + \boldsymbol{\delta}_1, \quad \text{with } \operatorname{rank}(\mathbf{W}_{1j}) = r_{1j} \ll d_j \text{ and } \mathbb{E}\|\boldsymbol{\delta}_1\|^2 \to 0.$$

Substituting into $\mathcal{T} = \mathbb{E}[\mathbf{Z}_1 \otimes \mathbf{Z}_2 \otimes \cdots \otimes \mathbf{Z}_M]$ and taking the mode-1 unfolding yields

$$\mathcal{T}_{(1)} = \tilde{\mathcal{T}}_{(1)} + \mathcal{E}_{(1)}, \quad \text{where } \tilde{\mathcal{T}}_{(1)} = \sum_{j=2}^{M} \mathbf{W}_{1j}\, \mathbb{E}\big[\mathbf{Z}_j(\mathbf{Z}_2 \otimes \cdots \otimes \mathbf{Z}_M)^\top\big].$$

By construction, $\operatorname{rank}(\tilde{\mathcal{T}}_{(1)}) \leq \sum_{j=2}^{M} r_{1j}$; let $R = \sum_{j=2}^{M} r_{1j}$. For the modality-wise normalized embeddings, Cauchy–Schwarz gives $\|\mathcal{E}_{(1)}\|_F \leq (\mathbb{E}\|\boldsymbol{\delta}_1\|^2)^{1/2} \to 0$. For any tolerance $\epsilon > \|\mathcal{E}_{(1)}\|_F$, $\tilde{\mathcal{T}}_{(1)}$ is a feasible rank-R approximation to $\mathcal{T}_{(1)}$. By the Eckart–Young–Mirsky theorem [12, 31], the optimal rank-R approximation error satisfies

$$\min_{\operatorname{rank}(A) \leq R} \|\mathcal{T}_{(1)} - A\|_F \leq \|\mathcal{E}_{(1)}\|_F < \epsilon.$$

Consequently, at most R singular values of $\mathcal{T}_{(1)}$ can exceed ϵ, *i.e.*,

$$\operatorname{rank}^{(\epsilon)}_{(1)}(\mathcal{T}) \leq R.$$

Consider the full orthogonal Tucker decomposition of $\mathcal{T}$ obtained via the Higher-Order Singular Value Decomposition (HOSVD):

$$\mathcal{T} = \mathcal{G}^{\text{full}} \times_1 Q^{(1)} \times_2 \cdots \times_M Q^{(M)},$$

where each $Q^{(m)} \in \mathbb{R}^{d_m \times d_m}$ is an orthogonal matrix. The mode-1 unfolding gives

$$\mathcal{T}_{(1)} = Q^{(1)} \mathcal{G}^{\text{full}}_{(1)} (Q^{(M)} \otimes \cdots \otimes Q^{(2)})^\top,$$

and by orthogonality, $\sigma_k(\mathcal{T}_{(1)}) = \sigma_k(\mathcal{G}^{\text{full}}_{(1)})$ for all k. Hence,

$$\operatorname{rank}^{(\epsilon)}_{(1)}(\mathcal{G}^{\text{full}}) = \operatorname{rank}^{(\epsilon)}_{(1)}(\mathcal{T}) \leq R.$$

The Gaussian-QR projection in Eq. (4) and the orthogonal Tucker factors used in Appendix B cannot increase the rank or residual norm of this approximation. Hence the truncated core also satisfies $\operatorname{rank}^{(\epsilon)}_{(1)}(\mathcal{G}) \leq R$.

□

Lemma 2 (Energy Concentration Indicates Degenerate Structure)**.** *Let*

$$\rho(\mathcal{G}) := \max_{i_1,\ldots,i_M} \frac{\mathcal{G}^2_{i_1,\ldots,i_M}}{\|\mathcal{G}\|_F^2}.$$

If $\rho(\mathcal{G}) \geq 1 - \delta$ for some $\delta \in [0, 1]$, then $\mathcal{G}$ is within relative Frobenius error at most $\sqrt{\delta}$ of a one-sparse core tensor. For small δ, this indicates dominance by a single separable component in the Tucker core.

Proof. For $\mathcal{G} \neq 0$, let $(i_1^\star, \ldots, i_M^\star)$ be an index attaining the maximum entry magnitude of $\mathcal{G}$, and define a one-sparse tensor $\mathcal{G}^{(1)}$ that keeps only $\mathcal{G}_{i_1^\star,\ldots,i_M^\star}$. Then

$$\|\mathcal{G} - \mathcal{G}^{(1)}\|_F^2 = \|\mathcal{G}\|_F^2 - \max_{i_1,\ldots,i_M} \mathcal{G}_{i_1,\ldots,i_M}^2.$$

If $\rho(\mathcal{G}) \geq 1 - \delta$, then

$$\frac{\|\mathcal{G} - \mathcal{G}^{(1)}\|_F^2}{\|\mathcal{G}\|_F^2} \leq \delta,$$

and hence

$$\frac{\|\mathcal{G} - \mathcal{G}^{(1)}\|_F}{\|\mathcal{G}\|_F} \leq \sqrt{\delta}.$$

Therefore $\mathcal{G}$ is close to a one-sparse tensor in relative Frobenius norm, which indicates a degenerate structure dominated by a single separable component when δ is small. □

Lemma 3 (Johnson–Lindenstrauss lemma [10])**.** *If the columns of $P^{(m)}$ are i.i.d. standard Gaussian vectors that have been orthonormalized, then for any fixed finite sample set $\{Z_b^{(m)}\}_{b=1}^B$ independent of $P^{(m)}$, by choosing $r_p = O(\epsilon^{-2} \log B)$, the following holds with high probability:*

$$\begin{aligned}(1-\epsilon)\frac{r_p}{d_m}\|Z_b^{(m)} - Z_{b'}^{(m)}\|_2^2 &\leq \|(Z_b^{(m)} - Z_{b'}^{(m)})P^{(m)}\|_2^2 \\ &\leq (1+\epsilon)\frac{r_p}{d_m}\|Z_b^{(m)} - Z_{b'}^{(m)}\|_2^2.\end{aligned}$$

In other words, pairwise distances are approximately preserved up to a common scale under the projection.

Proof. The lemma follows from [10], and we do not provide the proof here for brevity. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We focus on the synergistic information in self-supervised multimodal representation learning. We first characterizes the synergistic gap for more than two modalities, and then propose HRIL to preserve the information capacity of synergy. Extensive experiments on diverse real-world datasets verify the effectiveness of our method.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: As shown in Appendix A, we discuss the limitations of this work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems, formulas, propositions, and proofs in the main text are properly numbered and cross-referenced. The key proofs have been included in the appendix G.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Detailed analysis and results can be found in Sec. 4. Experimental details can be found in Appendix B. After the anonymity period, we will open-source our training code and key hyperparameters for reproducing.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: After the anonymity period, we will open-source our training code and key hyperparameters for reproducing.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: As shown in Appendix B, we provide comprehensive details of training procedures. Detailed implementation code will be released upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For all experiments, we report the mean and standard deviation across five independent runs with random seeds in the range $[42, 46]$.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We use a single NVIDIA RTX 4090 GPU with 24GB of VRAM for all experiments. More details can be found in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All experiments utilize publicly available benchmark datasets in accordance with their respective licensing terms. Since this work involves neither human subjects nor personally identifiable information, institutional review board (IRB) approval is not required. Comprehensive experimental configurations are documented in Appendix B.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to Appendix E.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We utilize the pre-processed datasets provided by MultiBench [27], which have been anonymized to safeguard personal privacy.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have provided detailed descriptions and clearly marked the sources and citations for all models and frameworks involved in the paper within the experimental section. For open-source code, we have included comprehensive comments and explanations for all imported packages and foundational code used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This work introduces new assets, including the training code and dataset developed for the HRIL representation learning framework. These resources are comprehensively documented in the experimental section, which details the model architecture, training protocols, and dataset specifications, alongside complete citations for all underlying frameworks and libraries. The accompanying open-source codebase features extensive inline annotations that clarify the functionality of external dependencies and the organization of core modules. A standardized documentation template is provided alongside the release, specifying training configurations, confirming dataset licensing compliance, and outlining model limitations.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: In this work, the LLM is utilized solely for writing and editing purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.