SYNCO: LEARNING CROSS-MODAL SYNERGY BY CONTRASTING INTERACTION RESIDUALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal contrastive learning is a dominant paradigm for learning transferable representations from unlabeled data, but standard objectives primarily capture information that is redundant between modalities. Partial Information Decomposition (PID) shows that task-relevant information in multimodal data decomposes into three components: redundancy shared between modalities, uniqueness specific to each modality, and synergy available only from their joint observation. Recent frameworks extend contrastive learning to capture all three components, yet synergy remains undertrained in practice. We propose SynCo (Synergy Contrastive Learning), a method that directly addresses synergy undertraining through dedicated supervision on an interaction residual. SynCo fits a linear projector to predict the fused representation from independently computed unimodal features, and the resulting interaction residual, which removes the linearly unimodal-predictable component, receives dedicated contrastive supervision at negligible computational cost. On the controlled Trifeature benchmark, SynCo achieves state-of-the-art synergy capture with a $+5.98\%$ gain over the baseline, and on real-world benchmarks from MultiBench, DARai, and MM-IMDb, SynCo consistently outperforms or matches prior methods across diverse modality combinations and task types. The method operates as a plug-in to existing contrastive multimodal frameworks without modifying the underlying fusion architecture and can further improve synergy capture when combined with other methods.

1 INTRODUCTION

Multimodal learning plays a central role in high-stakes applications such as medical imaging (Li et al., 2020; Prabhushankar et al., 2022), autonomous systems (Kokilepersaud et al., 2023), video understanding (Kaviani et al., 2025), and scientific applications (Orlyanchik et al., 2025; Massey et al., 2025). Such domains benefit from leveraging complementary cues across modalities, producing richer representations than unimodal approaches. Complementary signals are particularly valuable in real-world settings where data variation is significant (Hendrycks et al., 2021; Yarici et al., 2025). However, capturing the cross-modal interactions that produce these representations remains a fundamental challenge in multimodal learning.

Multimodal contrastive learning has emerged as the dominant approach to this challenge. Methods such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn joint representations across modalities using contrastive objectives (Oord et al., 2018; Chen et al., 2020). Contrastive methods achieve strong transfer across downstream tasks. However, these methods are built on the multi-view redundancy assumption, which holds that all task-relevant information is shared between modalities (Tsai et al., 2021; Tian et al., 2020a). As a result, they capture only the information shared between modalities and miss the complementary modality-specific information (Sridharan & Kakade, 2008). The assumption is restrictive, since many real-world tasks depend on complementary information unique to individual modalities (Baltrušaitis et al., 2018; Zong et al., 2024).

Partial Information Decomposition (PID) (Williams & Beer, 2010; Bertschinger et al., 2014) provides a principled framework for characterizing the limitations of redundancy-based objectives. For two modalities X_1 and X_2 predicting a task Y, PID decomposes the total mutual information $I(X_1, X_2; Y)$ into three components: redundancy (R), information shared across modalities; uniqueness (U_1, U_2), information specific to each modality; and synergy (S), information that emerges only

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from their joint observation. Each component carries distinct task-relevant information, and capturing all components is necessary for learning representations that reflect the full structure of multimodal information.

Despite recent advances in multimodal contrastive learning, effectively capturing synergistic information remains challenging. FactorCL (Liang et al., 2023b) explicitly separates shared and unique representations but relies on restrictive conditional augmentation assumptions and does not explicitly model synergy. CoMM (Dufumier et al., 2024) extends multimodal contrastive learning to capture redundancy, uniqueness, and synergy by maximizing mutual information between augmented multimodal representations alongside unimodal contrastive objectives. While redundancy and uniqueness are accessible from individual modalities, synergy requires their joint observation. Consequently, synergy can remain undertrained even when the contrastive objective theoretically preserves all three components. InfMasking (Wen et al., 2025) addresses this challenge by stochastically masking modality features during fusion and aligning masked representations with unmasked ones through mutual information maximization. COrAL (Cissee et al., 2026) promotes synergistic interactions through asymmetric masking while separating shared and modality-specific representations using orthogonality constraints. Both approaches encourage cross-modal interactions through masking-based training, without explicitly separating the linearly unimodal-predictable component from the fused representation.

We propose **SynCo** (**Syn**ergy **Co**ntrastive Learning) to address synergy undertraining in multimodal contrastive learning through dedicated supervision on the interaction residual. SynCo introduces a Synergy Head consisting of a linear projector that predicts the fused representation from unimodal representations and a residual operation that removes the linearly unimodal-predictable component. The resulting interaction residual is used to encourage synergy capture through a dedicated contrastive loss that complements the base objective. SynCo adds only two loss terms and a lightweight prediction head to the base framework, with negligible computational overhead in the evaluated setting and no modifications to the underlying fusion architecture.

- We introduce a residual decomposition that removes the linearly unimodal-predictable component of multimodal representations. A linear projector captures the part of the fused representation that is linearly predictable from unimodal features, and the resulting interaction residual provides a direct target for synergy-focused supervision.
- We propose SynCo, a method that applies dedicated contrastive supervision on the interaction residual to enhance synergy capture. SynCo operates as a plug-in to existing multimodal contrastive frameworks and adds only two loss terms and a lightweight prediction head at negligible computational overhead.
- We demonstrate SynCo's effectiveness across multiple benchmarks. On the Trifeature benchmark, SynCo achieves state-of-the-art synergy capture, exceeding the baseline by $+5.98\%$. On real-world benchmarks from MultiBench, DARai, and MM-IMDb, SynCo consistently outperforms or matches prior methods across vision-language, sensor-based, and clinical domains.

2 Related Work

Multimodal learning. Multimodal learning combines information from heterogeneous sources such as text, images, audio, and tabular data into a single model (Baltrušaitis et al., 2018; Liang et al., 2024). Early methods combined modalities either at the input level through feature concatenation or at the decision level by aggregating unimodal predictions (Snoek et al., 2005; Baltrušaitis et al., 2018). More recent work uses transformer architectures to model cross-modal interactions directly through attention (Lu et al., 2019; Xu et al., 2023), enabling applications across tasks such as translation (Xu et al., 2015; Rombach et al., 2022) and representation learning (Bengio et al., 2013). Among these directions, self-supervised representation learning has emerged as a leading approach for learning general-purpose representations from unlabeled data through contrastive objectives (Zong et al., 2024; Radford et al., 2021; Jia et al., 2021). Our work targets this setting.

Self-supervised multimodal representation learning. Self-supervised learning trains models using objectives defined directly on the input data, without requiring human labels (Balestriero et al., 2023). In the multimodal setting, these methods use cross-modal information to learn transferable representations (Zong et al., 2024). Prior work can be organized by the type of pretext task. Clustering-based approaches use cluster assignments as pseudo-labels, allowing modalities

to supervise one another (Alwassel et al., 2020). Masked modeling methods reconstruct masked portions of the input, often in a cross-modal fashion where one modality predicts missing information from another (Bachmann et al., 2022; Mizrahi et al., 2023). Contrastive methods align representations of matched multimodal pairs while pushing unmatched pairs apart (Radford et al., 2021; Jia et al., 2021). Our work builds on this contrastive line.

Contrastive multimodal interactions. Standard contrastive multimodal methods (Radford et al., 2021; Jia et al., 2021) optimize cross-modal alignment but primarily capture redundant information shared across modalities. FactorCL (Liang et al., 2023b) explicitly factorizes shared and unique information, but relies on restrictive conditional augmentation assumptions and does not model synergy. CoMM (Dufumier et al., 2024) grounds its objectives in Partial Information Decomposition (Williams & Beer, 2010; Bertschinger et al., 2014), combining multimodal and unimodal contrastive losses to capture redundancy, uniqueness, and synergy. In practice, however, synergy remains the most difficult component to learn, since it requires cross-modal computation. InfMasking (Wen et al., 2025) addresses this by stochastically masking modality features during fusion, which encourages cross-modal computation through input perturbation but does not supervise synergy directly. COrAL (Cissee et al., 2026) separates shared and modality-specific features through a dual-path architecture with orthogonality constraints and applies asymmetric masking with complementary patterns across views, which also promotes synergy through input perturbation. In contrast, our method targets synergy explicitly through a dedicated contrastive objective on the interaction residual.

3 PRELIMINARIES

We introduce the problem setup and Partial Information Decomposition, and review the multimodal contrastive learning framework introduced by CoMM (Dufumier et al., 2024).

Problem setup. Let $X_1, X_2, \ldots, X_n$ be random variables representing n data modalities and let Y denote a target variable for a downstream task. The objective is to learn a multimodal representation $Z = f_\theta(X)$, where $X = (X_1, \ldots, X_n)$ and θ denotes the parameters of a multimodal encoder, such that Z preserves all task-relevant information: $I(Z;Y) = I(X;Y)$. During pretraining, two augmentations t' and t'' are drawn from a set $\mathcal{T}$ of multimodal augmentations, independently of each other and of (X, Y), and applied to the same input. We write $X' = t'(X)$ and $X'' = t''(X)$ for the two views and $Z' = f_\theta(X')$ and $Z'' = f_\theta(X'')$ for their representations. The two views are identically distributed and differ only in the realized augmentation. For each modality i, the projection t_i masks all modalities except the i-th, and $Z_i = f_\theta(t_i(X))$ denotes the corresponding unimodal representation, while $Z_i' = f_\theta(t_i(X'))$ and $Z_i'' = f_\theta(t_i(X''))$ denote its counterparts for the two views. Since Z and every Z_i are outputs of the same encoder f_θ, all representations lie in the same space $\mathbb{R}^D$, where D is the output dimension of f_θ.

Partial Information Decomposition. To characterize the structure of task-relevant information, we employ Partial Information Decomposition (PID) (Williams & Beer, 2010; Bertschinger et al., 2014; Liang et al., 2023a). Our theoretical analysis is presented for $n = 2$ modalities, since multimodal interactions have not been characterized by PID for larger n (Dufumier et al., 2024). For two modalities X_1 and X_2 and a target Y, PID decomposes the total task-relevant information $I(X_1, X_2; Y)$ into four non-negative terms:

$$I(X_1, X_2; Y) = R + S + U_1 + U_2, \tag{1}$$

where R denotes redundancy (information shared across modalities), U_i denotes uniqueness (information specific to modality i), and S denotes synergy (information available only from the joint observation of both modalities). This decomposition satisfies the consistency equations:

$$I(X_1; Y) = R + U_1, \quad I(X_2; Y) = R + U_2, \quad I(X_1; X_2; Y) = R - S, \tag{2}$$

Multimodal contrastive learning. In self-supervised learning, Y remains unspecified during training, so task-relevant information cannot be optimized directly. CoMM (Dufumier et al., 2024) addresses this by assuming the existence of multimodal augmentations that preserve all task-relevant information, allowing it to be optimized indirectly through contrastive objectives between augmented views.

Assumption 1 (Minimal label-preserving multimodal augmentations, Dufumier et al., 2024)**.** The set $\mathcal{T}$ of multimodal augmentations preserves task-relevant content: for any $t \in \mathcal{T}$ and $X' = t(X)$, $I(X; X') = I(X; Y)$ and $I(X'; Y) = I(X; Y)$.

The assumption states that augmenting X to X' preserves the information X carries about the target Y, even though Y itself is not used to construct the augmentation. Unlike standard contrastive assumptions, the preserved information is not required to be shared across modalities. The set $\mathcal{T}$ therefore includes augmentations that are jointly defined on (X_1, X_2) rather than applied independently to each modality. Joint augmentations can remove content from one modality when the other modality retains it, which keeps $I(X'; Y) = I(X; Y)$. The first equality is the condition stated in CoMM, and the second equality states explicitly that the augmentation preserves the information about Y. Appendix E discusses Assumption 1 further.

Let $X' = t(X)$ for some $t \in \mathcal{T}$, and let $Z = f_\theta(X)$ and $Z' = f_\theta(X')$ denote the representations of the original and augmented inputs. Since the pair (Z, Z') is a deterministic function of (X, X'), the data processing inequality yields $I(Z; Z') \leq I(X; X')$. Combined with Assumption 1, this leads to the following result:

Lemma 1 (Multimodal information preservation, Dufumier et al., 2024)**.** *Under sufficient expressivity of f_θ, the optimal parameters $\theta^\star$ that maximize $I(Z_\theta; Z'_\theta)$ satisfy $I(Z_{\theta^\star}; Z'_{\theta^\star}) = I(X; X') = I(X; Y)$ and $I(Z_{\theta^\star}; Y) = I(X; Y)$.*

Since $I(X; Y) = I(X_1, X_2; Y) = R + S + U_1 + U_2$ by Eq. (1), Lemma 1 implies that maximizing mutual information between augmented multimodal representations captures all PID components. The two augmented views X' and X'' satisfy the same conditions, and the representation at the optimum preserves the task-relevant information of X (Lemma 5, Appendix D). A corresponding result holds for the unimodal case when a single modality is used as the input. Under Assumption 1, the augmented view preserves the information that modality i carries about Y, and the unimodal contrastive terms reach this quantity at the optimum.

Lemma 2 (Unimodal information preservation, Dufumier et al., 2024)**.** *Let $\theta^\star$ be the optimal parameters satisfying Lemma 1, which under sufficient expressivity of f_θ also maximize $I(Z_i; Z')$. For the projection t_i that selects modality i, the corresponding unimodal representation $Z_i = f_{\theta^\star}(t_i(X))$ satisfies*

$$I(Z_i; Z') = I(Z_i; Y) = I(X_i; Y) = R + U_i. \tag{3}$$

The equalities follow from $I(X_i; X') = I(X_i; Y)$ under Assumption 1 and the data processing inequality applied to the chains $Y \to X_i \to Z_i$ and $Z_i \to Y \to Z'$, together with the sufficient expressivity of f_θ on the unimodal input. The same result holds with Z'' in place of Z' (Appendix D).

CoMM (Dufumier et al., 2024) introduces Lemmas 1 and 2 and uses them to learn task-agnostic representations that capture redundancy, uniqueness, and synergy. Mutual information is estimated using the InfoNCE lower bound (Oord et al., 2018) with temperature τ, where $\hat{I}_{\text{NCE}}(A, B)$ denotes the InfoNCE estimate of $I(A; B)$, giving the multimodal contrastive objective:

$$\mathcal{L}_{\text{MCL}} = -\hat{I}_{\text{NCE}}(Z', Z'') - \tfrac{1}{2} \sum_{i=1}^{n} \left[\hat{I}_{\text{NCE}}(Z_i, Z') + \hat{I}_{\text{NCE}}(Z_i, Z'') \right]. \tag{4}$$

At the theoretical optimum, the first term captures $R + S + \sum_i U_i$, and the second term captures $\sum_i (R + U_i)$ in the bimodal setting (Lemmas 1 and 2). Although the first term theoretically captures synergy, learning synergistic interactions remains challenging in practice because they require cross-modal computation.

4 Synergy Contrastive Learning

4.1 The SynCo Framework

We propose SynCo (Synergy Contrastive Learning), a framework that addresses synergy undertraining by introducing dedicated supervision for synergistic interactions in fused multimodal representations. SynCo introduces a Synergy Head consisting of a linear projector P followed by a residual operation. The projector predicts the fused representation from the unimodal representations, and the component

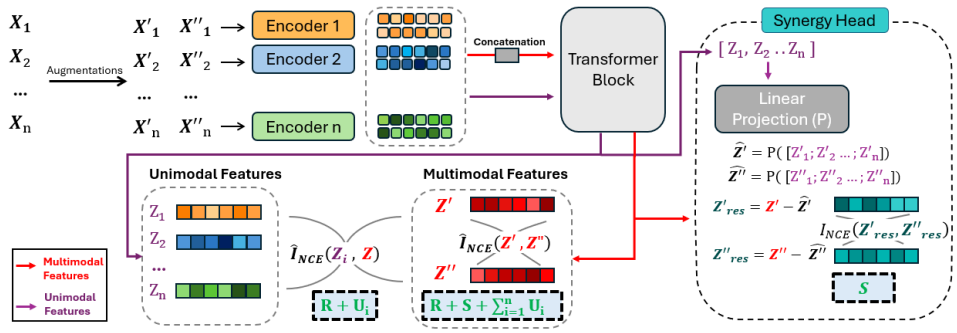


Figure 1: Overview of SynCo. Two augmented views are encoded and processed through a multimodal fusion path producing Z' and Z'', and a unimodal path producing $Z_1, \ldots, Z_n$. Three contrastive objectives target distinct PID components: $\hat{I}_{\text{NCE}}(Z', Z'')$ captures $R + S + \sum_i U_i$, $\hat{I}_{\text{NCE}}(Z_i, Z')$ and $\hat{I}_{\text{NCE}}(Z_i, Z'')$ capture $R + U_i$, and $\hat{I}_{\text{NCE}}(Z'_{\text{res}}, Z''_{\text{res}})$ provides synergy-focused supervision on the interaction residuals.

of the fused representation that the projector cannot predict forms the interaction residual. The interaction residual removes the linearly unimodal-predictable component while retaining information not captured by the linear prediction. A contrastive loss is applied to the residual to encourage synergy capture. The overall pipeline is illustrated in Figure 1.

SynCo extends the multimodal contrastive objective $\mathcal{L}_{\text{MCL}}$ of Eq. (4) with two additional terms that operate on the Synergy Head: a contrastive loss $\mathcal{L}_{\text{interaction}}$ on the interaction residuals from the two augmented views, which provides synergy-focused supervision, and a mean-squared-error loss $\mathcal{L}_{\text{pred}}$ that trains the linear projector P. The complete SynCo training objective is:

$$\mathcal{L}_{\text{SynCo}} = \mathcal{L}_{\text{MCL}} + \underbrace{\lambda_{\text{int}}\,\mathcal{L}_{\text{interaction}} + \lambda_{\text{pred}}\,\mathcal{L}_{\text{pred}}}_{\text{synergy-targeted terms}}, \tag{5}$$

where λ_{int} and λ_{pred} are the weighting coefficients. The theoretical foundation supporting this claim is presented in Section 4.2.

4.2 Theoretical Foundation

We show that the residual $Z^\star_{\text{res}}$ provides a target for synergy-focused supervision by removing the linearly unimodal-predictable content while retaining additional task-relevant information given the linear prediction. The argument combines an empirically supported hypothesis about the trained encoder with an information-theoretic decomposition.

Let $P : \mathbb{R}^{2D} \to \mathbb{R}^D$ be a linear projector trained by mean-squared error to predict the fused multimodal representation Z from the unimodal representations $[Z_1; Z_2]$. We denote the prediction by $\hat{Z} = P([Z_1; Z_2])$ and define the residual $Z_{\text{res}} = Z - \hat{Z}$ as the component of Z not captured by the prediction. We write $P^\star$ for the MSE-optimal projector and $\hat{Z}^\star$, $Z^\star_{\text{res}}$ for the corresponding prediction and residual.

Hypothesis 1 (Synergy is not linearly recoverable from unimodal representations)**.** For the trained encoder f_θ and the MSE-optimal projector $P^\star$, the linear prediction $\hat{Z}^\star = P^\star([Z_1; Z_2])$ satisfies $I(\hat{Z}^\star; Y) \leq R + U_1 + U_2$.

The hypothesis bounds the task-relevant information retained by the linear prediction of the fused representation from unimodal representations. We assess it empirically with linear probes in Section 5.1.

To combine this hypothesis with the structure of mutual information, we first establish how task-relevant information is distributed between the prediction $\hat{Z}$ and the residual Z_{res}.

Lemma 3 (Information decomposition of the residual)**.** *For any encoder f_θ, any $\hat{Z} = P([Z_1; Z_2])$, and $Z_{\text{res}} = Z - \hat{Z}$,*

$$I(Z_{\text{res}}; Y \mid \hat{Z}) \geq I(Z; Y) - I(\hat{Z}; Y), \tag{6}$$

with equality when $I(Z;Y) = I(X;Y)$.

Proof. Given $\hat{Z}$, the residual determines Z via $Z = \hat{Z} + Z_{\text{res}}$, so $I(Z_{\text{res}}; Y \mid \hat{Z}) = I(Z; Y \mid \hat{Z})$. The chain rule gives $I(Z, \hat{Z}; Y) = I(\hat{Z}; Y) + I(Z; Y \mid \hat{Z})$, and $I(Z, \hat{Z}; Y) \geq I(Z; Y)$ yields the inequality. Both Z and $\hat{Z}$ are deterministic functions of X, so by the data processing inequality $I(Z, \hat{Z}; Y) \leq I(X; Y)$. When $I(Z; Y) = I(X; Y)$, the two bounds give $I(Z, \hat{Z}; Y) = I(Z; Y)$, and the inequality holds with equality. □

Theorem 1 (Residual concentrates synergistic information)**.** *Let $Z = f_\theta(X_1, X_2)$ for any encoder f_θ, and let $\Delta = I(X;Y) - I(Z;Y) \geq 0$ denote the task-relevant information that Z does not capture. If Hypothesis 1 holds for f_θ, then the residual $Z^\star_{\text{res}} = Z - \hat{Z}^\star$ satisfies*

$$I(Z^\star_{\text{res}}; Y \mid \hat{Z}^\star) \geq S - \Delta. \tag{7}$$

In particular, when $\Delta = 0$, the residual satisfies $I(Z^\star_{\text{res}}; Y \mid \hat{Z}^\star) \geq S$, providing at least S additional task-relevant information given the linear prediction.

Proof. By Lemma 3 and Hypothesis 1, $I(Z^\star_{\text{res}}; Y \mid \hat{Z}^\star) \geq I(Z;Y) - I(\hat{Z}^\star; Y) \geq I(Z;Y) - (R + U_1 + U_2)$. Substituting $I(Z;Y) = I(X;Y) - \Delta = R + S + U_1 + U_2 - \Delta$ from Eq. (1) gives the lower bound, which reduces to S when $\Delta = 0$. □

Remark 1. Theorem 1 holds for any encoder that satisfies Hypothesis 1. At the optimum of $\mathcal{L}_{\text{MCL}}$, Lemma 1 and its extension to two augmented views (Lemma 5) give $\Delta = 0$. The linear prediction then carries at most $R + U_1 + U_2$ of the task-relevant information, while the residual provides at least S additional task-relevant information given the linear prediction. We refer to this conditional information bound as concentration of synergistic information. The linear projector removes the linearly unimodal-predictable component of the fused representation, while nonlinear $R + U$ content beyond the reach of the projector may remain in $Z^\star_{\text{res}}$.

Corollary 1 (Synergy-targeted contrastive learning)**.** *Let Z'_{res} and Z''_{res} be the residuals from two augmented views of the same sample. Under Assumption 1 and Hypothesis 1, the contrastive objective $\hat{I}_{\text{NCE}}(Z'_{\text{res}}, Z''_{\text{res}})$ rewards task-relevant information shared between the interaction residuals, providing a basis for synergy-focused supervision of the multimodal representation.*

Proof. Assumption 1 preserves the total task-relevant information $I(X;Y)$ across both augmented views. Theorem 1 applied to each view, with $\Delta = 0$ at the information-preserving optimum (Lemma 5), shows that each interaction residual provides at least S additional task-relevant information given its linear prediction, which captures the linearly unimodal-predictable component of the fused representation. The InfoNCE objective $\hat{I}_{\text{NCE}}(Z'_{\text{res}}, Z''_{\text{res}})$ is a lower bound on $I(Z'_{\text{res}}; Z''_{\text{res}})$ and rewards information shared between the two residuals. Since the two views are conditionally independent given Y (Lemma 5), the residuals are also conditionally independent given Y. The data processing inequality therefore gives

$$I(Z'_{\text{res}}; Z''_{\text{res}}) \leq \min\left\{I(Z'_{\text{res}}; Y), I(Z''_{\text{res}}; Y)\right\}. \tag{8}$$

Thus, the information shared between the interaction residuals is task-relevant under Assumption 1. Applying InfoNCE to these residuals encourages the encoder to preserve shared task-relevant information in the component remaining after subtraction of the linear prediction, providing the motivation for synergy-focused supervision. □

4.3 Implementation

We apply the theoretical setup of Section 4.2 to two augmented views, where D denotes the dimensionality of the fusion output and n the number of modalities. The MSE objective from Section 4.2 is written as

$$\mathcal{L}_{\text{pred}} = \frac{1}{2}\left[\|Z' - P([Z'_1; Z'_2])\|^2 + \|Z'' - P([Z''_1; Z''_2])\|^2\right]. \tag{9}$$

The interaction residual is obtained by subtracting P's prediction from the fused representation:

$$Z'_{\text{res}} = Z' - \hat{Z}', \quad Z''_{\text{res}} = Z'' - \hat{Z}'', \quad \text{where} \quad \hat{Z}' = P([Z'_1; Z'_2]),\ \hat{Z}'' = P([Z''_1; Z''_2]). \tag{10}$$

The interaction contrastive loss applies InfoNCE to ℓ_2-normalized residuals from the two augmented views, which form a positive pair targeted at the synergistic component (Corollary 1):

$$\mathcal{L}_{\text{interaction}} = -\hat{I}_{\text{NCE}}\left(\frac{Z'_{\text{res}}}{\|Z'_{\text{res}}\|_2}, \frac{Z''_{\text{res}}}{\|Z''_{\text{res}}\|_2}\right), \tag{11}$$

using the same temperature τ as $\mathcal{L}_{\text{MCL}}$. Across all experiments, we use fixed weights $\lambda_{\text{int}} = 0.01$ and $\lambda_{\text{pred}} = 0.1$ without dataset-specific tuning, and linearly warm up λ_{int} over the first 20 epochs to its full value. The sensitivity analyses in Appendices B.3 and B.6 justify these defaults. Full implementation details are provided in Appendix A.

Gradient isolation. Stop-gradient operators separate the parameter updates of P from those of the encoder. In $\mathcal{L}_{\text{pred}}$, we detach both the target (Z', Z'') and the unimodal inputs $(Z'_1, Z'_2, Z''_1, Z''_2)$ to P, restricting this loss to update only the parameters of P. In $\mathcal{L}_{\text{interaction}}$, we detach the predictions $\hat{Z}'$ and $\hat{Z}''$ before forming the residuals, restricting gradient propagation to Z' and Z'' and the multimodal fusion path of the encoder, with no flow into P or the unimodal representations. This construction maintains the convergence of P to the MSE-optimal projector $P^\star$ required by Theorem 1 and directs gradient updates to the component of the encoder responsible for cross-modal computation.

5 Experiments

We evaluate SynCo on controlled synthetic experiments and large-scale real-world benchmarks. We compare against cross-modal contrastive learning (Cross) (Radford et al., 2021), its extension with modality-wise self-supervised terms (Cross+Self) (Yuan et al., 2021), the disentanglement-based FactorCL (Liang et al., 2023b), the multimodal interaction framework CoMM (Dufumier et al., 2024), and the masking-based InfMasking (Wen et al., 2025) and COrAL (Cissee et al., 2026). We also add the SynCo terms to InfMasking for additional comparison.

On a synthetic bimodal dataset derived from Trifeature (Hermann & Lampinen, 2020), we measure how well the learned representations capture each PID component: redundancy, uniqueness, and synergy. We then assess generalization on real-world tasks from MultiBench (Liang et al., 2021), DARai (Kaviani et al., 2025), and MM-IMDb (Arevalo et al., 2017). These benchmarks span diverse domains and modality combinations. We evaluate all methods via linear probing: after self-supervised pretraining, we freeze the encoder and train a linear classifier or regressor on the learned representations. We report mean and standard deviation over 5 independent runs. Appendix A provides full experimental details.

5.1 Controlled Experiments on the Trifeature Dataset

We evaluate SynCo on a controlled synthetic benchmark following the experimental protocol of CoMM (Dufumier et al., 2024). The dataset is derived from Trifeature (Hermann & Lampinen, 2020), where each image is defined by three visual attributes: shape, texture, and color, each taking one of ten possible values. Bimodal inputs are constructed by pairing two images. This construction allows precise control over which attributes are shared or unique to each modality and which cross-modal relationships are synergistic. We construct probing tasks for each PID component. The two constructions below define different pretraining sets, and we pretrain a separate model for each. For redundancy and uniqueness, we pair images by matching shape, making shape the redundant attribute and texture the unique attribute of each modality. We then evaluate linear probes predicting the shared shape (redundancy) and the texture of the first image (uniqueness), both with a chance level of 10%. For synergy, we define a fixed bijection $\mathcal{M}$ between the ten textures and the ten colors. The self-supervised pretraining set consists of pairs (X_1, X_2) satisfying $\mathcal{M}$, which establishes a cross-modal relationship between $\text{texture}(X_1)$ and $\text{color}(X_2)$ that cannot be recovered from either modality in isolation. The probing task is binary classification of whether a given pair satisfies $\mathcal{M}$, with label $Y = \mathbf{1}[(\text{texture}(X_1), \text{color}(X_2)) \in \mathcal{M}]$. The probe training and test sets are class-balanced, with positive pairs drawn from the pretraining distribution and negative pairs constructed from texture-color combinations outside $\mathcal{M}$, yielding a chance level of 50%.

Results are reported in Table 1. Cross-modal contrastive training alone (Cross) achieves perfect redundancy but fails entirely on both uniqueness and synergy. Adding per-modality self-supervised terms (Cross+Self) recovers uniqueness but fails to improve synergy beyond chance. FactorCL

Table 1: Linear probing accuracy (%) for redundancy, uniqueness, and synergy on Trifeature. Bold marks the best value per section. †: results from Dufumier et al. (2024).

Model	*redundancy*↑	*uniqueness*↑	*synergy*↑
Cross†	**100.0**	11.6	50.0
Cross+Self†	99.7	86.9	50.0
FactorCL†	99.8	62.5	46.5
COrAL	$99.7_{\pm 0.06}$	$\mathbf{91.4}_{\pm 0.62}$	$73.91_{\pm 0.51}$
CoMM	$99.9_{\pm 0.05}$	$86.6_{\pm 1.32}$	$71.6_{\pm 1.07}$
SynCo (ours)	$99.9_{\pm 0.21}$	$90.15_{\pm 1.14}$	$\mathbf{77.58}_{\pm 0.74}$
InfMasking	$\mathbf{99.9}_{\pm 0.07}$	$89.1_{\pm 0.95}$	$76.12_{\pm 1.42}$
SynCo (ours) + InfMasking	$\mathbf{99.9}_{\pm 0.24}$	$\mathbf{90.51}_{\pm 1.27}$	$\mathbf{77.89}_{\pm 0.85}$

Table 2: Linear probing accuracy (%) on the unimodal representations and on the three components of SynCo's representation decomposition across PID tasks on the Trifeature benchmark. $[Z_1; Z_2]$: unimodal; Z: multimodal; $\hat{Z}$: linear prediction; Z_{res}: residual.

Representation	*Redun.*↑	*Uniq.*↑	*Syn.*↑
$[Z_1; Z_2]$	$99.9_{\pm 0.16}$	$89.42_{\pm 2.33}$	$50.00_{\pm 0.00}$
Z	$99.9_{\pm 0.21}$	$90.15_{\pm 1.14}$	$77.58_{\pm 0.74}$
$\hat{Z}$	$99.9_{\pm 0.18}$	$89.26_{\pm 2.41}$	$50.00_{\pm 0.00}$
Z_{res}	$81.1_{\pm 2.17}$	$57.40_{\pm 2.25}$	$\mathbf{87.24}_{\pm 1.78}$

performs below chance on synergy, confirming that disentanglement-based objectives do not resolve the synergy deficit. CoMM is the first method to capture all three PID components, and InfMasking further improves synergy through stochastic feature masking. SynCo surpasses CoMM by $+3.55\%$ on uniqueness and $+5.98\%$ on synergy while maintaining near-perfect redundancy, and outperforms InfMasking by $+1.46\%$ on synergy despite using a substantially simpler mechanism. COrAL achieves the highest uniqueness, while SynCo exceeds it by $+3.67\%$ on synergy. Combining SynCo with InfMasking yields the highest synergy among all methods. The consistent gains over CoMM and InfMasking confirm that our method provides effective synergy-focused supervision of the learned representations.

Linear probing on decomposed representations. We assess Hypothesis 1 empirically by training linear probes on the concatenated unimodal representations $[Z_1; Z_2]$ and on the three components of SynCo's representation: the multimodal representation Z, the linear prediction $\hat{Z} = P([Z_1; Z_2])$, and the residual $Z_{\text{res}} = Z - \hat{Z}$. Table 2 reports the results. Neither the concatenated unimodal representations nor the linear prediction $\hat{Z}$ exceeds chance-level accuracy on the synergy task, indicating that the synergy label is not linearly decodable from the unimodal representations or from their linear prediction, consistent with Hypothesis 1. The residual Z_{res} surpasses the multimodal representation Z on synergy by $+9.66\%$, consistent with Theorem 1. The residual also contains nonzero R and U content, consistent with Remark 1, since the linear projector removes only the linearly unimodal-predictable component and nonlinear $R + U$ content may remain in the residual.

5.2 Experiments on Real-world Datasets

We evaluate SynCo on real-world multimodal datasets from MultiBench (Liang et al., 2021), DARai (Kaviani et al., 2025), and MM-IMDb (Arevalo et al., 2017). These benchmarks span diverse domains, modality types, and task formulations. Following prior work (Liang et al., 2023b; Dufumier et al., 2024; Wen et al., 2025), we use the same data preprocessing steps, modality configurations, and backbone networks across all methods for MultiBench and MM-IMDb. For DARai, we adopt the sensor encoder from (Khaertdinov et al., 2021). Implementation details are provided in Appendix A.

MultiBench. We use five datasets from MultiBench: Vision&Touch (Lee et al., 2020) for end-effector position regression from visual and proprioceptive modalities, MIMIC (Johnson et al., 2016) for binary classification of respiratory-system diagnoses from tabular and time-series modalities, and MOSI (Zadeh et al., 2016), UR-FUNNY (Hasan et al., 2019), and MUSTARD (Castro et al., 2019)

Table 3: Linear probing on MultiBench: MSE ($\times 10^{-4}$) for regression and top-1 accuracy (%) for classification. Bold: best within each section. †: results from Dufumier et al. (2024). *: classification average.

Model	*Regr.*	*Classification*				
	V&T↓	*MIMIC*↑	*MOSI*↑	*UR-FUNNY*↑	*MUSTARD*↑	**Avg.***↑
Cross†	$33.09_{\pm 3.67}$	$66.70_{\pm 0.10}$	$47.80_{\pm 1.80}$	$50.10_{\pm 1.90}$	$53.50_{\pm 2.90}$	54.53
Cross+Self†	$7.56_{\pm 0.31}$	$65.49_{\pm 0.00}$	$49.00_{\pm 1.10}$	$59.90_{\pm 0.90}$	$53.90_{\pm 4.00}$	57.07
FactorCL†	$10.82_{\pm 0.56}$	$67.30_{\pm 0.00}$	$51.20_{\pm 1.60}$	$60.50_{\pm 0.80}$	$55.80_{\pm 0.90}$	58.70
COrAL	$5.40_{\pm 0.28}$	$\mathbf{67.95}_{\pm 0.21}$	$66.40_{\pm 0.84}$	$64.80_{\pm 0.47}$	$66.10_{\pm 1.02}$	66.31
CoMM	$6.85_{\pm 2.13}$	$65.92_{\pm 0.35}$	$63.70_{\pm 2.50}$	$63.25_{\pm 0.81}$	$64.58_{\pm 1.50}$	64.36
SynCo (ours)	$\mathbf{5.21}_{\pm 0.34}$	$67.89_{\pm 0.56}$	$\mathbf{67.92}_{\pm 1.06}$	$\mathbf{66.52}_{\pm 0.53}$	$\mathbf{67.02}_{\pm 1.12}$	**67.34**
InfMasking	$\mathbf{4.95}_{\pm 0.56}$	$67.53_{\pm 0.32}$	$67.12_{\pm 0.75}$	$64.11_{\pm 0.95}$	$65.62_{\pm 0.90}$	66.10
SynCo + InfMasking	$5.12_{\pm 0.21}$	$\mathbf{69.52}_{\pm 0.49}$	$\mathbf{67.81}_{\pm 1.18}$	$\mathbf{66.91}_{\pm 0.62}$	$\mathbf{67.21}_{\pm 1.65}$	**67.86**

for sentiment analysis, humor detection, and sarcasm detection from visual and textual modalities. Appendix G describes each dataset in detail. Table 3 reports the results. SynCo outperforms CoMM, InfMasking, and COrAL on average classification accuracy and achieves the highest accuracy on MOSI among standalone methods. Combining SynCo with InfMasking yields the best overall performance, with the highest accuracy on MIMIC, UR-FUNNY, and MUSTARD, and the strongest average across classification tasks. On the regression task (V&T), SynCo remains competitive with InfMasking, outperforms COrAL, and substantially outperforms all other baselines.

We report results on the wearable-sensor benchmark DARai (Kaviani et al., 2025) in Appendix B.1 and results in the trimodal setting in Appendix B.2.

MM-IMDb. MM-IMDb (Arevalo et al., 2017) is a multi-label movie genre classification dataset with visual (poster) and textual (plot summary) modalities. The dataset is challenging due to substantial class imbalance across genres and a large semantic gap between the two modalities. We compare against CLIP (Radford et al., 2021), SLIP (Mu et al., 2022), CoMM (Dufumier et al., 2024), and InfMasking (Wen et al., 2025), with all methods using a CLIP backbone. Table 4 reports the results. SynCo outperforms both CoMM and standalone InfMasking on weighted-F1 and macro-F1 and achieves the best standalone performance. Combining SynCo with InfMasking yields a further gain and the best overall result on both metrics.

Table 4: Linear probing F1-scores (weighted and macro, in %) on MM-IMDb (vision and language). All self-supervised methods use a CLIP backbone. † denotes results from Dufumier et al. (2024).

Model	*w-F1*↑	*m-F1*↑
CLIP† Radford et al. (2021)	$54.49_{\pm 0.19}$	$44.94_{\pm 0.30}$
SLIP† Mu et al. (2022)	$56.54_{\pm 0.19}$	$47.35_{\pm 0.27}$
CoMM Dufumier et al. (2024)	$61.29_{\pm 0.73}$	$53.79_{\pm 0.22}$
SynCo (ours)	$\mathbf{62.12}_{\pm 0.17}$	$\mathbf{55.41}_{\pm 0.16}$
InfMasking Wen et al. (2025)	$61.97_{\pm 0.28}$	$55.28_{\pm 0.15}$
SynCo + InfMask.	$\mathbf{62.23}_{\pm 0.27}$	$\mathbf{55.64}_{\pm 0.22}$

5.3 Ablation Studies

Ablation on Projector Architecture. The residual $Z_{\text{res}} = Z - \hat{Z}$ contains the part of Z that P cannot predict, so a higher-capacity projector may capture more of the fused representation, leaving less information in the residual. To assess this design choice, we replace P with five higher-capacity variants: 2-layer, 3-layer, and 5-layer MLPs, a Gated Residual projector, and a 4-layer Transformer. Holding all other components fixed, we evaluate linear probing accuracy on the Trifeature synergy task for Z, $\hat{Z}$, and Z_{res}. Architectural details are provided in Appendix A.7.

Table 5: Synergy probe accuracy (%) on Trifeature across projector architectures. Chance level is 50%.

Projector	$Z\uparrow$	$\hat{Z}\downarrow$	$Z_{\text{res}}\uparrow$
Linear (ours)	**77.58**	**50.00**	**87.24**
MLP-2	74.14	59.75	75.31
MLP-3	73.11	62.10	74.13
MLP-5	71.25	64.81	72.84
Gated Residual	70.56	68.19	69.41
Transformer-4	66.45	71.47	62.04

Table 5 shows that the prediction $\hat{Z}$ of the linear projector remains at chance on the synergy task, while higher-capacity projectors increasingly capture information useful for the synergy task in $\hat{Z}$. As projector capacity increases, synergy probe accuracy decreases in both Z_{res} and Z, weakening the effectiveness of $\mathcal{L}_{\text{interaction}}$ for synergy-focused supervision. The linear projector achieves the highest synergy probe accuracy in Z_{res} while removing the linearly unimodal-predictable component of the fused representation. Additional ablations on the loss terms, loss weights, and warmup length are provided in Appendices B.4, B.3, and B.6, respectively.

6 Conclusion

We presented SynCo, a contrastive multimodal framework that addresses synergy undertraining by supervising an interaction residual that removes the linearly unimodal-predictable component of the fused representation. The Synergy Head provides synergy-focused supervision at negligible computational cost. On the controlled Trifeature benchmark, SynCo achieves state-of-the-art synergy capture and improves uniqueness while maintaining near-perfect redundancy. On real-world benchmarks from MultiBench, DARai, and MM-IMDb, SynCo consistently outperforms or matches prior methods across diverse modality combinations and task types, including vision-language, sensor-based, and clinical domains. SynCo adds only two loss terms and a lightweight prediction head without modifying the underlying fusion architecture, making it a straightforward plug-in for existing multimodal contrastive frameworks.

AI USE STATEMENT

We used large language models in the preparation of this work. For tasks with required disclosure, we used large language models to provide feedback on the research methodology and experiments, to assist in revising the paper content, and to assist in implementation coding. We did not use large language models to generate synthetic data sets, to develop the theoretical framework, to formulate the mathematical claims or provide critical ingredients for their proofs, to propose or refine the research hypothesis, or to interpret the experimental results, and the remaining tasks with required disclosure, namely translation, dataset cleaning and reformatting, and qualitative or thematic data analysis, are not applicable to this work. We also used large language models for retrieval and discovery, to identify relevant literature. All AI-assisted work was reviewed by the authors, the implementation was verified and tested by the authors, and all reported numbers come from experiments run by the authors. We take responsibility for the final content of this work, including the text, the claims, and the artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The encoder architectures, augmentations, optimization settings, and linear probing protocol for every benchmark are described in Appendix A, and the datasets, splits, and preprocessing steps are described in Appendix G. SynCo uses the same coefficients $\lambda_{\text{int}} = 0.01$ and $\lambda_{\text{pred}} = 0.1$ and the same 20-epoch warmup across all experiments without dataset-specific tuning, and the sensitivity of SynCo to the loss weights and the warmup length is reported in Appendices B.3 and B.6. The assumptions underlying the theoretical results are stated in Section 3 and Section 4.2, and complete proofs of all lemmas, the theorem, and the corollary are given in Appendix D. The projector variants used in the architecture ablation are specified in Appendix A.7, and the computational cost of the method is reported in Appendix C. All experiments run on a single NVIDIA A40 GPU, and all reported results are averaged over five independent runs. We will publicly release the implementation of SynCo, including the training scripts and configuration files for every experiment, upon publication.

REFERENCES

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in neural information processing systems*, 33:9758–9770, 2020.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European conference on computer vision*, pp. 348–367. Springer, 2022.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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

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*, pp. 4619–4629, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Carolin Cissee, Raneen Younis, and Zahra Ahmadi. Orthogonalized multimodal contrastive learning with asymmetric masking for structured representations. *arXiv preprint arXiv:2602.14983*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? *arXiv preprint arXiv:2409.07402*, 2024.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 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)*, pp. 2046–2056, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

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

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*, pp. 4904–4916. PMLR, 2021.

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.

Ghazal Kaviani, Yavuz Yarici, Seulgi Kim, Mohit Prabhushankar, Ghassan AlRegib, Mashhour Solh, and Ameya Patil. Hierarchical and multimodal data for daily activity understanding. *Journal of Data-centric Machine Learning Research (DMLR)*, 2025.

Bulat Khaertdinov, Esam Ghaleb, and Stylianos Asteriadis. Contrastive self-supervised learning for sensor-based human activity recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–8. IEEE, 2021.

Kiran Kokilepersaud, Mohit Prabhushankar, Yavuz Yarici, Ghassan AlRegib, and Armin Parchami. Exploiting the distortion-semantic interaction in fisheye data. *IEEE Open Journal of Signal Processing*, 4:284–293, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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.

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning. *CoRR*, abs/2107.07502, 2021. URL `https://arxiv.org/abs/2107.07502`.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023a.

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

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM computing surveys*, 56(10): 1–42, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Matthew Massey, Nusrat Munia, and Abdullah-Al-Zubaer Imran. Earthscape: A multimodal dataset for surficial geologic mapping and earth surface analysis. *arXiv preprint arXiv:2503.15625*, 2025.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pp. 529–544. Springer, 2022.

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

Vladimir Orlyanchik, Jeffrey Kimmell, Zackary Snow, Amir Ziabari, and Vincent Paquit. A co-registered in-situ and ex-situ dataset from wire arc additive manufacturing process. *Scientific Data*, 12(1):343, 2025.

Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pp. 3982–3992, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 399–402, 2005.

Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In Rocco A. Servedio and Tong Zhang (eds.), *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pp. 403–414. Omnipress, 2008. URL http://colt2008.cs.helsinki.fi/papers/94-Sridharan.pdf.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020a.

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*, 33:6827–6839, 2020b.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective, 2021. URL https://arxiv.org/abs/2006.05576.

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.

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

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

Yavuz Yarici, Kiran Kokilepersaud, Mohit Prabhushankar, and Ghassan AlRegib. Subject invariant contrastive learning for human activity recognition. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2025.

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*, pp. 6995–7004, 2021.

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.

Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299–5318, 2024.

A Appendix Implementation Details

This section describes the encoder architectures, modality-specific augmentations, latent converters, and optimization settings used across all experiments.

A.1 Encoder Architectures

Trifeature. Both modalities are visual, so we use a shared AlexNet backbone (Krizhevsky et al., 2012) with a 512-dimensional embedding space, matching the encoder choice of CoMM. We remove the final average-pooling layer and apply a linear patch-embedding module (Dosovitskiy et al., 2020) to the resulting 6×6 feature maps. Fixed two-dimensional sine-cosine positional embeddings are added before the fusion transformer.

MIMIC. The two modalities of MIMIC are tabular and time-series. The tabular branch uses a two-layer MLP with a 10-dimensional hidden size and a 10-dimensional output, and the time-series branch uses a GRU with a 512-dimensional hidden state. A feature tokenizer (Gorishniy et al., 2021) maps the tabular output to a 512-dimensional embedding space, and no latent converter is applied to the time-series branch.

MOSI, UR-FUNNY, and MUSTARD. For these three datasets the modalities are pre-extracted visual and textual features. We follow FactorCL (Liang et al., 2023b) and CoMM in using a five-head, five-layer Transformer with a 40-dimensional embedding space for each modality, with no latent converter.

Vision&Touch. For the end-effector regression task we use the visual and force-torque modalities. Images are encoded with a ResNet-18 (He et al., 2016) producing 128-channel spatial feature maps, and force-torque signals are encoded with a five-layer causal convolutional network producing a 128-dimensional output, following the original Vision&Touch protocol (Lee et al., 2020). A patch-embedding module is applied to the ResNet-18 feature maps, and the force-torque embeddings are tokenized through a feature tokenizer.

DARai. For each wearable modality we adopt the encoder architecture introduced for sensor-based human activity recognition by Khaertdinov et al. (Khaertdinov et al., 2021), which combines a stack of three one-dimensional convolutional layers, each followed by batch normalization and a ReLU activation and using reflective padding to preserve the temporal length, with a transformer self-attention block. We fix the kernel size to three throughout the convolutional stack and use eight attention heads in the transformer block. All sensor streams are resampled to a common length of 100 time steps before encoding, so that the IMU, EMG, and insole branches operate on tensors of identical temporal extent. The three branches share this backbone configuration and differ only in the channel count of the input layer, which is determined by the corresponding sensor dimensionality.

MM-IMDb. The visual encoder is a ViT-B/32 (Dosovitskiy et al., 2020) pre-trained with CLIP (Radford et al., 2021), and the textual encoder is the multilingual Sentence-BERT model distilled from CLIP (Reimers & Gurevych, 2019). Both encoders are frozen during pretraining, and SynCo operates directly on the token embeddings produced by these backbones, without latent converters.

A.2 Fusion Transformer and Linear Projector

The fusion module is a Transformer encoder layer with multi-head self-attention, a feed-forward block, residual connections, and layer normalization. We use a single layer with eight heads in the bimodal setting and two layers with eight heads in the trimodal setting. A learnable `[CLS]` token is prepended to the input sequence, and the corresponding output token is read out as the fused representation Z. The unimodal representation Z_i is the `[CLS]` output of the same modality encoders and fusion transformer when all modalities except the i-th are masked, so the fused and unimodal representations lie in the same space $\mathbb{R}^D$. The linear projector P in SynCo is implemented as a single linear layer mapping $\mathbb{R}^{nD}$ to $\mathbb{R}^D$, where D is the dimensionality of Z and n is the number of modalities. It is applied to the concatenation of the unimodal representations $[Z_1; \ldots; Z_n]$ and produces $\hat{Z}$.

A.3 Residual Computation and Training Dynamics

Gradient structure. The three stop-gradient placements introduced in Section 4.3 yield a clear separation of learning roles. The linear projector P receives gradients only from $\mathcal{L}_{\text{pred}}$ and learns to approximate Z from $[Z_1; Z_2]$ to the extent permitted by a linear function. The encoder and fusion transformer receive gradients from two distinct sources: the multimodal contrastive loss $\mathcal{L}_{\text{MCL}}$, which propagates through the projection head and operates in the projected space, and the interaction contrastive loss $\mathcal{L}_{\text{interaction}}$, which propagates through Z'_{res} and Z''_{res} and operates in the raw fusion space. The interaction contrastive loss specifically supervises features that P cannot predict, which by Theorem 1 provide at least S additional task-relevant information given the linear prediction. The construction does not introduce conflicting optimization objectives. P is trained to minimize prediction error and therefore cannot increase the residual norm. The encoder receives no gradient that would degrade the unimodal representations to enlarge the residual, since the predictions $\hat{Z}'$ and $\hat{Z}''$ are detached in $\mathcal{L}_{\text{interaction}}$ and no gradient from this loss passes through P or the unimodal representations. The modality encoders and fusion transformer are shared between the fused and unimodal paths, so updates driven by $\mathcal{L}_{\text{interaction}}$ still change the unimodal representations indirectly, and the unimodal terms of $\mathcal{L}_{\text{MCL}}$ keep them informative about $R + U_i$ (Lemma 2).

Training dynamics. SynCo exhibits a natural curriculum arising from the interplay of prediction and interaction objectives, and the implementation additionally performs an explicit linear warmup of the interaction weight. Early in training the projector P is randomly initialized and the residual approximates the fused representation, so the interaction contrastive term initially provides a broad discriminative signal. To avoid the interaction term dominating too early, we linearly scale the effective interaction weight from zero to its configured value over a fixed number of warmup epochs:

$$\lambda_{\text{int}}^{\text{eff}} = \lambda_{\text{int}} \cdot \min\left(1, \frac{\text{epoch}}{\text{warmup_epochs}}\right).$$

In our runs, `warmup_epochs` is set to 20, so λ_{int} is fully active only after the warmup period. The prediction loss $\mathcal{L}_{\text{pred}}$ is applied from the start, while the interaction loss is gradually annealed in as P converges, so that the residual progressively excludes the linearly unimodal-predictable component. Beyond this linear warmup of λ_{int}, no additional scheduling or curriculum is required.

A.4 Modality-Specific Augmentations

For raw images on Trifeature, MM-IMDb, and Vision&Touch, we adopt the SimCLR augmentation pipeline (Chen et al., 2020). Tabular features in MIMIC are perturbed with additive Gaussian noise applied independently to each feature dimension. For pre-extracted time-series features in MIMIC, MOSI, UR-FUNNY, MUSTARD, and the force-torque modality of Vision&Touch, we apply Gaussian noise together with random dropping of a uniformly sampled fraction of the sequence between 0% and 80%; dropped timesteps are zeroed rather than removed. For the wearable signals in DARai, we do not use sequence dropping; instead we apply sensor-level augmentations including additive Gaussian jittering, optional scaling, random sign and rotation, permutation of segments, and channel shuffling, as implemented in the DARai transforms. For raw text on MM-IMDb, we apply random token masking with probability 15%, where tokens are selected uniformly with pad, CLS, and SEP tokens excluded, by replacing selected token ids with the tokenizer's mask token. This is a direct mask-replacement scheme applied only during training and does not implement the BERT $80/10/10$ replace/keep schedule.

A.5 Optimization

All experiments use the AdamW optimizer (Loshchilov & Hutter, 2017), except for DARai where we use Adam following the protocol of the source benchmark. The learning rate is set to 3×10^{-4} with weight decay 10^{-4} on Trifeature, 1×10^{-3} with weight decay 10^{-2} on MIMIC, MOSI, UR-FUNNY, and MUSTARD, 1×10^{-4} with weight decay 10^{-2} on Vision&Touch and MM-IMDb, and 1×10^{-3} on DARai. Pretraining uses a batch size of 64 across all datasets. For MM-IMDb we additionally use a cosine schedule with a final value of 10^{-6} and a ten-epoch warmup. Pretraining is run for 100 epochs on every dataset except MM-IMDb, where we run 70 epochs following CoMM, and DARai, where we pretrain for 300 epochs following the source benchmark. The InfoNCE critic

is implemented as a three-layer MLP with 512-dimensional hidden layers and a 256-dimensional output, matching the projection head architecture of SimCLR. The temperature τ in every InfoNCE term is set to 0.1 and shared across $\mathcal{L}_{\text{MCL}}$ and $\mathcal{L}_{\text{interaction}}$. Across all experiments we use the fixed coefficients $\lambda_{\text{int}} = 0.01$ and $\lambda_{\text{pred}} = 0.1$ for the two SynCo-specific terms. The sensitivity analysis in Appendix B.3 justifies these defaults. All experiments are performed on a single NVIDIA A40 GPU with 48 GB of memory.

A.6 Linear Probing Protocol

After pretraining, the encoder and fusion transformer are frozen, and a single linear layer is trained on top of the multimodal representation Z for 100 epochs. Classification tasks are optimized with cross-entropy loss, and the Vision&Touch end-effector task is optimized with mean-squared error. We use early stopping on the validation set and report mean and standard deviation across five independent runs.

A.7 Projector Architecture Ablation Details

This appendix details the projector variants evaluated in the ablation reported in Section 5.3. All variants implement the mapping $P : \mathbb{R}^{nD} \to \mathbb{R}^D$ from the concatenated unimodal representations $[Z_1; \ldots; Z_n]$ to the fused representation space, where D is the dimensionality of Z and n is the number of modalities. Throughout the ablation, only the architecture of P is varied; the encoder, fusion transformer, augmentations, optimizer, loss coefficients, and training schedule are held fixed at the values reported in Appendix A, and each variant is trained jointly with the SynCo objective. The stop-gradient configuration from Section 4 continues to apply, so $\mathcal{L}_{\text{pred}}$ updates only the parameters of P regardless of its architecture. All variants are described under the bimodal Trifeature setting with $n = 2$ and $D = 512$.

Linear (baseline). A single fully connected layer with weight matrix $W \in \mathbb{R}^{D \times nD}$ and bias $b \in \mathbb{R}^D$, giving $\hat{Z} = W[Z_1; Z_2] + b$. This is the projector used in all main-text results.

2-layer MLP. A two-layer feed-forward network. The first layer maps $\mathbb{R}^{nD}$ to a hidden dimension $h = (nD + D)/2$, followed by a ReLU activation. The second layer maps the hidden representation to $\mathbb{R}^D$.

3-layer and 5-layer MLPs. Stacks of three and five fully connected layers respectively, with ReLU activations between consecutive layers and constant intermediate widths equal to the output dimension D. The first layer maps $\mathbb{R}^{nD}$ to $\mathbb{R}^D$ and all subsequent layers operate within $\mathbb{R}^D$.

Gated Residual. A linear branch combined additively with a parallel nonlinear branch:

$$\hat{Z} = W_\ell[Z_1; Z_2] + b_\ell + \gamma \cdot \text{MLP}([Z_1; Z_2]), \tag{12}$$

where $W_\ell \in \mathbb{R}^{D \times nD}$ and $b_\ell \in \mathbb{R}^D$ parameterize the linear branch, MLP is a two-layer feed-forward network with hidden dimension D and ReLU activation, and $\gamma \in \mathbb{R}$ is a learnable scalar gate initialized to zero. Initializing γ to zero renders the projector strictly linear at the start of training, so the residual loss begins from the same starting point as the linear baseline, and γ grows only when the nonlinear branch reduces $\mathcal{L}_{\text{pred}}$.

Transformer (4-layer). The concatenated input $[Z_1; Z_2] \in \mathbb{R}^{nD}$ is reshaped into n modality tokens of dimension D, and learned modality-specific embeddings $E_1, \ldots, E_n \in \mathbb{R}^D$ are added to the corresponding tokens. The resulting sequence is processed by a stack of four Transformer encoder layers, each containing multi-head self-attention with eight heads, a position-wise feed-forward sublayer with hidden dimension $4D$ and GELU activation, and pre-layer normalization on both sublayers. The output tokens are flattened to $\mathbb{R}^{nD}$ and passed through a two-layer MLP head with hidden dimension D and GELU activation, producing $\hat{Z} \in \mathbb{R}^D$. This variant comprises approximately 14.7M parameters.

A.8 COMBINING SYNCO WITH INFMASKING

Combining SynCo with InfMasking augments the multimodal contrastive terms with a masking objective $\mathcal{L}_{\text{mask}}$ that stochastically replaces fusion-input tokens with a learned mask token and maximizes agreement between the masked and unmasked multimodal representations. Its total loss is the uniform mean of the masking terms and the contrastive terms of $\mathcal{L}_{\text{MCL}}$; we denote it $\mathcal{L}_{\text{InfMasking}}$ and leave it untouched. The combined objective is

$$\mathcal{L}_{\text{SynCo+InfMasking}} = \mathcal{L}_{\text{InfMasking}} + \lambda_{\text{int}}\,\mathcal{L}_{\text{interaction}} + \lambda_{\text{pred}}\,\mathcal{L}_{\text{pred}}. \tag{13}$$

The Synergy Head operates on the unmasked representations. The residuals Z'_{res} and Z''_{res} are formed from the fused representations Z' and Z'' computed without InfMasking's feature masking, and the projector P receives the unimodal representations Z'_i and Z''_i defined in Section 3. The stochastically masked representations enter only $\mathcal{L}_{\text{mask}}$. InfMasking perturbs the input to fusion, while SynCo decomposes the unmasked output of fusion, and the two mechanisms act on different quantities.

The stop-gradient configuration of Section 4.3 is unchanged, and $\mathcal{L}_{\text{mask}}$ does not update P. We use the InfMasking hyperparameters of Wen et al. (2025) for the masking ratio, the number of masked views, and λ_{mask}, together with the fixed SynCo coefficients $\lambda_{\text{int}} = 0.01$ and $\lambda_{\text{pred}} = 0.1$ and the 20-epoch warmup of λ_{int}. All other settings follow Appendix A. The combination requires no forward passes beyond those of InfMasking, and its only additional cost is the linear projector.

B APPENDIX ADDITIONAL EXPERIMENTS

B.1 EXPERIMENTS ON DARAI

DARai (Kaviani et al., 2025) is a human activity recognition dataset with multiple wearable sensor modalities, and we use it to evaluate whether SynCo's gains generalize across modality pairings. We evaluate three bimodal configurations, IMU + Insole, IMU + EMG, and EMG + Insole, and compare FactorCL, CoMM, InfMasking, and SynCo under identical preprocessing and backbones. Appendix G describes the sensor modalities and the three configurations in detail.

Table 6: Linear probing top-1 accuracy (%) on DARai across three bimodal sensor configurations. Bold values indicate the best within each section.

Model	*IMU+Insole*↑	*IMU+EMG*↑	*EMG+Insole*↑	**Avg.**↑
FactorCL	$36.74_{\pm 1.15}$	$34.18_{\pm 1.05}$	$30.92_{\pm 1.18}$	33.95
CoMM	$42.12_{\pm 0.81}$	$41.85_{\pm 0.94}$	$32.21_{\pm 1.05}$	38.73
SynCo (ours)	$\mathbf{45.52}_{\pm 0.54}$	$\mathbf{43.78}_{\pm 0.42}$	$\mathbf{36.15}_{\pm 0.59}$	**41.82**
InfMasking	$43.85_{\pm 1.13}$	$42.32_{\pm 0.84}$	$36.04_{\pm 1.12}$	40.74
SynCo (ours) + InfMasking	$\mathbf{45.21}_{\pm 0.98}$	$\mathbf{44.15}_{\pm 1.07}$	$\mathbf{36.38}_{\pm 1.06}$	**41.91**

Table 6 reports the results. SynCo outperforms CoMM on every configuration and achieves the best standalone average. Combining SynCo with InfMasking yields an additional gain on average, with the best mean accuracy on IMU + EMG and EMG + Insole, while standalone SynCo leads on IMU + Insole. The consistent gains across configurations show that SynCo generalizes beyond vision-language settings to wearable sensor data.

B.2 EXPERIMENTS WITH 3 MODALITIES

We extend our evaluation to the trimodal setting using MOSI, UR-FUNNY, and MUSTARD with their full text, audio, and video configurations. While Section 5.2 reports results on the bimodal variants of these datasets, the trimodal setting introduces additional cross-modal interactions that test whether the synergy-targeted supervision in SynCo scales beyond two modalities. We follow the same experimental protocol as the bimodal evaluation, comparing against CoMM and InfMasking under identical preprocessing, backbone networks, and linear probing procedures. Table 7 reports the results. SynCo achieves the strongest performance across all three datasets and on average, outperforming CoMM by $+2.94\%$ and InfMasking by $+1.35\%$ in average classification accuracy. The improvement over CoMM is consistent with the bimodal results and indicates that the residual

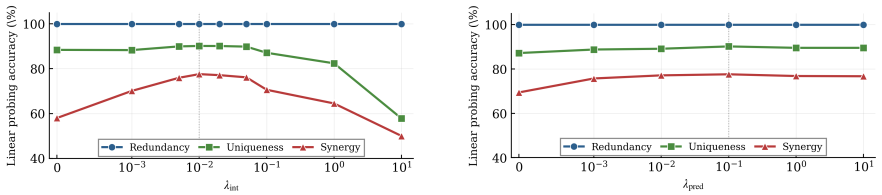
Table 7: Linear probing top-1 accuracy (in %) on trimodal classification benchmarks (text, audio, video).

Model	*MOSI*↑	*UR-FUNNY*↑	*MUSTARD*↑	**Average**↑
CoMM Dufumier et al. (2024)	$64.00_{\pm 2.33}$	$64.74_{\pm 0.71}$	$65.52_{\pm 1.42}$	64.75
SynCo (ours)	$\mathbf{68.11}_{\pm 0.92}$	$\mathbf{66.85}_{\pm 0.55}$	$\mathbf{68.12}_{\pm 1.24}$	**67.69**
InfMasking Wen et al. (2025)	$67.15_{\pm 1.21}$	$64.95_{\pm 0.92}$	$66.92_{\pm 2.75}$	66.34
SynCo (ours) + InfMasking	$\mathbf{67.52}_{\pm 1.11}$	$\mathbf{66.18}_{\pm 0.57}$	$\mathbf{67.48}_{\pm 1.48}$	**67.06**

decomposition continues to provide effective synergy-focused supervision as the number of modalities increases. Combining SynCo with InfMasking improves over standalone InfMasking on average but does not exceed standalone SynCo in the trimodal setting.

B.3 Ablation Experiment: Sensitivity to Loss Weights

The two coefficients in the SynCo objective control disjoint parameter sets due to the stop-gradient configuration: λ_{pred} governs only the optimization of P, while λ_{int} scales the encoder gradient from $\mathcal{L}_{\text{interaction}}$. We sweep each coefficient on Trifeature with the other held at its default value.



(a) λ_{int} sweep (b) λ_{pred} sweep

Figure 2: Sensitivity of SynCo's loss weights on Trifeature. (a) $\lambda_{\text{pred}} = 0.1$ fixed. (b) $\lambda_{\text{int}} = 0.01$ fixed. Dotted lines mark the values used for all experiments.

Figure 2a reports the sweep over λ_{int} with $\lambda_{\text{pred}} = 0.1$. Synergy rises with λ_{int}, reaches a peak at $\lambda_{\text{int}} = 0.01$, and degrades at larger values. Redundancy remains stable across all values, while uniqueness degrades alongside synergy at very large λ_{int}, which indicates that overly large values disrupt overall representation quality. Figure 2b reports the sweep over λ_{pred} with $\lambda_{\text{int}} = 0.01$. Once λ_{pred} is sufficiently large to train P effectively, all three components remain approximately flat across several orders of magnitude. The flat response is consistent with the gradient isolation property: λ_{pred} scales the loss for P alone and does not directly affect the encoder. The two sweeps show that $\lambda_{\text{int}} = 0.01$ yields the best synergy performance and $\lambda_{\text{pred}} = 0.1$ provides a sufficient threshold for training P; both are used as fixed defaults throughout all experiments.

B.4 Ablation Experiment: Loss Terms

The SynCo objective in Eq. (5) adds two terms to $\mathcal{L}_{\text{MCL}}$, and the interaction residual is defined relative to the fused representation that the multimodal term of Eq. (4) shapes. To isolate the contribution of each component, we train three objectives on Trifeature under identical settings: $\mathcal{L}_{\text{MCL}}$ alone, the full SynCo objective, and the SynCo objective without the multimodal term $\hat{I}_{\text{NCE}}(Z', Z'')$. All other hyperparameters are held fixed at the values reported in Appendix A, and we report mean and standard deviation across five independent runs.

Table 8 reports the results. The residual term accounts for most of the gain over the baseline: without the multimodal term, SynCo still improves synergy by 3.90 points over $\mathcal{L}_{\text{MCL}}$, and adding the multimodal term recovers the remaining 2.08 points. The multimodal term therefore remains necessary. It drives Z toward $\Delta = 0$ in Theorem 1, and the residual is informative about synergy to the extent that the fused representation it is computed from preserves the task-relevant information. Uniqueness follows the same pattern, while redundancy is unaffected by either term.

Table 8: Linear probing accuracy (%) of redundancy, uniqueness, and synergy on Trifeature under different combinations of loss terms.

Objective	*redundancy*↑	*uniqueness*↑	*synergy*↑
$\mathcal{L}_{\text{MCL}}$ (CoMM)	$\mathbf{99.9}_{\pm 0.05}$	$86.6_{\pm 1.32}$	$71.6_{\pm 1.07}$
SynCo w/o $\hat{I}_{\text{NCE}}(Z', Z'')$	$99.7_{\pm 0.37}$	$88.1_{\pm 0.49}$	$75.5_{\pm 0.28}$
SynCo (full)	$\mathbf{99.9}_{\pm 0.21}$	$\mathbf{90.15}_{\pm 1.14}$	$\mathbf{77.58}_{\pm 0.74}$

B.5 Linear Probes on Unimodal Representations

Hypothesis 1 concerns the linear prediction of the fused representation from the unimodal representations, and Table 2 is consistent with it for the MSE-optimal projector $P^\star$. Here we probe the unimodal representations directly. We freeze the SynCo encoder trained on the Trifeature synergy task and train linear probes on the individual unimodal representations Z_1 and Z_2, on their concatenation $[Z_1; Z_2]$, and on the fused representation Z, following the protocol of Appendix A.

Table 9: Linear probing accuracy (%) on the Trifeature synergy task for the unimodal representations, their concatenation, and the fused representation. Chance level is 50%.

Representation	*synergy*↑
Z_1 (modality 1)	$50.00_{\pm 0.00}$
Z_2 (modality 2)	$50.00_{\pm 0.00}$
$[Z_1; Z_2]$ (concatenation)	$50.00_{\pm 0.00}$
Z (fused)	$\mathbf{77.58}_{\pm 0.74}$

Table 9 reports the results. Both unimodal representations and their concatenation remain at chance level, while the fused representation reaches 77.58%. The gap does not come from missing attribute information: texture and color are individually recoverable from the corresponding unimodal representation, and the synergy label depends on whether the particular texture-color pairing belongs to $\mathcal{M}$. Recovering the label therefore requires a function that combines the two modalities, which no linear function of $[Z_1; Z_2]$ provides. Since the linear prediction $\hat{Z} = P([Z_1; Z_2])$ is one such function, the concatenation result supports the premise of Hypothesis 1 independently of how well P is trained.

B.6 Ablation Experiment: Warmup Epoch Ablation

The interaction weight λ_{int} is linearly annealed from zero to its full value over a fixed number of warmup epochs at the start of training, as described in Appendix A. Early in training the linear projector P is randomly initialized, so the residual $Z_{\text{res}} = Z - \hat{Z}$ initially approximates the fused representation itself rather than the component that excludes linearly unimodal-predictable content. The warmup delays the residual loss until P has had time to approach its MSE-optimal solution. To assess the sensitivity of SynCo to this schedule, we sweep `warmup_epochs` on Trifeature while holding all other hyperparameters fixed at the values reported in Appendix A.

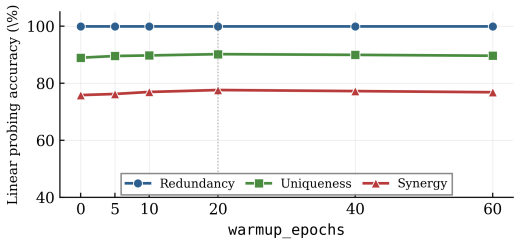


Figure 3: Sensitivity of SynCo to the warmup length on Trifeature. We sweep `warmup_epochs` $\in \{0, 5, 10, 20, 40, 60\}$ with all other hyperparameters fixed. Linear probing accuracy (%) is reported for each PID component, averaged over five independent runs. The dotted line marks the value used for all main-text experiments.

Figure 3 reports the results. All three PID components remain stable across the sweep, with redundancy unaffected and uniqueness and synergy varying within roughly two points across the full range. Performance peaks at 20 epochs and declines mildly at the endpoints. We use 20 as the default across all experiments.

C Appendix Processing Times and Complexity

We benchmark the computational cost of SynCo against CoMM and InfMasking to characterize the overhead introduced by the residual decomposition relative to existing multimodal contrastive frameworks. All measurements are obtained on a single GPU with batch size 64, averaged over multiple training steps. Table 10 reports per-step time, peak memory, and parameter counts for each method.

Table 10: Computational requirements of SynCo, CoMM, and InfMasking. Overhead is reported relative to CoMM. All measurements use batch size 64 on a single GPU.

Method	*Time/step (ms)*	*Memory (GB)*	*Parameters*	*Overhead*
CoMM	21.91 ± 0.23	0.17	448,464	baseline
SynCo (ours)	21.94 ± 0.18	0.17	451,704	$+0.15\%$
InfMasking	57.82 ± 0.33	0.71	448,504	$+163.92\%$

The results confirm that SynCo's residual decomposition introduces negligible computational cost. Relative to CoMM, SynCo increases per-step time by only 0.15% and adds $3{,}240$ parameters, corresponding to the linear projector that maps concatenated unimodal features to the fused representation space. Memory consumption is essentially unchanged. By contrast, InfMasking incurs substantially higher cost than both CoMM and SynCo, with per-step time more than doubling and peak memory increasing by approximately a factor of four. This overhead reflects the masking strategy of InfMasking, which requires multiple forward passes over masked feature combinations during training. SynCo achieves comparable or superior synergy capture to InfMasking on the Trifeature benchmark and competitive performance across MultiBench and MM-IMDb at a fraction of the computational cost, making it the more efficient choice when training resources are constrained. The overhead of SynCo is bounded by the size of the linear projector, which scales as $O(nD^2)$ in the number of modalities n and the fusion dimension D.

D Appendix Proofs

This appendix provides detailed proofs of the theoretical results stated in the main text. Assumption 1 and Lemmas 1 and 2 follow CoMM (Dufumier et al., 2024), and we prove the two lemmas in the notation of Section 3. Hypothesis 1 is assessed empirically in Section 5.1 and is treated as a premise rather than proven here. Throughout, the augmentations t, t', and t'' are drawn from $\mathcal{T}$ independently of each other and of (X, Y), so each augmented view depends on (X, Y) only through X and the chains $Y \to X \to X'$ and $Y \to X \to X''$ are Markov. We write sums for discrete variables; the arguments are identical with integrals for continuous variables.

Equivalent Form of Assumption 1

Lemma 4. *Let $X' = t(X)$ with $t \in \mathcal{T}$. The equalities $I(X; X') = I(X; Y)$ and $I(X'; Y) = I(X; Y)$ of Assumption 1 hold if and only if $X' \perp X \mid Y$ and $Y \perp X \mid X'$.*

Proof. Step 1: Label preservation. The Markov chain $Y \to X \to X'$ gives $I(X'; Y \mid X) = 0$, and the chain rule expands $I(X, X'; Y)$ in two orders as

$$I(X; Y) \;=\; I(X, X'; Y) \;=\; I(X'; Y) + I(X; Y \mid X'). \tag{14}$$

The equality $I(X'; Y) = I(X; Y)$ therefore holds if and only if $I(X; Y \mid X') = 0$, that is, $Y \perp X \mid X'$.

Step 2: Minimality. The chain rule expands $I(X; X', Y)$ in two orders as

$$I(X; X') + I(X; Y \mid X') \;=\; I(X; Y) + I(X; X' \mid Y). \tag{15}$$

When $I(X; Y \mid X') = 0$, Eq. (15) reduces to $I(X; X') - I(X; Y) = I(X; X' \mid Y)$, and the equality $I(X; X') = I(X; Y)$ holds if and only if $I(X; X' \mid Y) = 0$, that is, $X' \perp X \mid Y$. □

The condition $X' \perp X \mid Y$ states that the augmented view carries no information about X beyond the label, and the condition $Y \perp X \mid X'$ states that the view retains all information that X carries about the label.

PROOF OF LEMMA 1

Proof. Let $Z_\theta = f_\theta(X)$ and $Z'_\theta = f_\theta(X')$ with $X' = t(X)$ for some $t \in \mathcal{T}$.

Step 1: Upper bounds. The pair (Z_θ, Z'_θ) is a deterministic function of (X, X'), and the data processing inequality together with Assumption 1 gives

$$I(Z_\theta; Z'_\theta) \;\le\; I(X; X') \;=\; I(X; Y). \tag{16}$$

By Lemma 4, $X' \perp X \mid Y$. Since Z_θ and Z'_θ are functions of X and X' respectively, $Z'_\theta \perp Z_\theta \mid Y$, and the data processing inequality applied to the chains $Z_\theta \to Y \to Z'_\theta$ and $Y \to X \to Z_\theta$ gives

$$I(Z_\theta; Z'_\theta) \;\le\; I(Z_\theta; Y) \;\le\; I(X; Y). \tag{17}$$

Step 2: Tightness under sufficient expressivity. Under sufficient expressivity, the encoder family $\{f_\theta\}$ contains a map that is injective on the supports of X and X', such as the identity, and such a map attains the bound in Eq. (16). Hence

$$I(Z_{\theta^\star}; Z'_{\theta^\star}) \;=\; \max_\theta I(Z_\theta; Z'_\theta) \;=\; I(X; X') \;=\; I(X; Y). \tag{18}$$

Substituting Eq. (18) into Eq. (17) gives $I(X; Y) \;\le\; I(Z_{\theta^\star}; Y) \;\le\; I(X; Y)$, and therefore $I(Z_{\theta^\star}; Y) = I(X; Y)$. □

The first term of Eq. (4) contrasts two augmented views rather than an input and its augmentation. The following lemma shows that the pair (X', X'') satisfies the conditions used in the proof above.

Lemma 5. *Let $X' = t'(X)$ and $X'' = t''(X)$ with $t', t'' \in \mathcal{T}$. Under Assumption 1, $X' \perp X'' \mid Y$ and $I(X'; X'') = I(X'; Y) = I(X''; Y) = I(X; Y)$.*

Proof. Given X, the views X' and X'' are independent of each other and of Y. By Lemma 4, $X' \perp X \mid Y$, which together with the Markov chain $Y \to X \to X'$ gives $p(x' \mid x) = p(x' \mid y)$ for every (x, y) in the support. Therefore

$$p(x', x'' \mid y) \;=\; \sum_x p(x \mid y)\, p(x' \mid x)\, p(x'' \mid x) \;=\; p(x' \mid y)\, p(x'' \mid y), \tag{19}$$

which proves $X' \perp X'' \mid Y$. Lemma 4 applied to X'' gives $Y \perp X \mid X''$, and therefore

$$p(y, x' \mid x'') \;=\; \sum_x p(x \mid x'')\, p(y \mid x'')\, p(x' \mid x) \;=\; p(y \mid x'')\, p(x' \mid x''), \tag{20}$$

which proves $I(X'; Y \mid X'') = 0$. Expanding $I(X'; X'', Y)$ in two orders, as in Eq. (15), gives

$$I(X'; X'') \;=\; I(X'; Y) + I(X'; X'' \mid Y) - I(X'; Y \mid X'') \;=\; I(X'; Y), \tag{21}$$

and Assumption 1 gives $I(X'; Y) = I(X''; Y) = I(X; Y)$. □

With Lemma 5 in place of Lemma 4, the proof of Lemma 1 applies to Z'_θ and Z''_θ. The optimal parameters that maximize $I(Z'_\theta; Z''_\theta)$ therefore satisfy $I(Z'_{\theta^\star}; Z''_{\theta^\star}) = I(Z'_{\theta^\star}; Y) = I(Z''_{\theta^\star}; Y) = I(X; Y) = R + S + U_1 + U_2$, which is the quantity targeted by the first term of Eq. (4).

PROOF OF LEMMA 2

Proof. Step 1: Assumption 1 for a single modality. The selected modality $X_i = t_i(X)$ is a deterministic function of X. By Lemma 4, $X' \perp X \mid Y$ and $Y \perp X \mid X'$, which imply $X' \perp X_i \mid Y$ and $Y \perp X_i \mid X'$. Expanding $I(X_i; X', Y)$ in two orders, as in Eq. (15), gives

$$I(X_i; X') \;=\; I(X_i; Y) + I(X_i; X' \mid Y) - I(X_i; Y \mid X') \;=\; I(X_i; Y) \;=\; R + U_i, \tag{22}$$

where the last equality follows from the consistency equations in Eq. (2).

Step 2: Upper bounds. The unimodal representation $Z_i = f_\theta(t_i(X))$ is a function of X_i, and the data processing inequality together with Eq. (22) gives

$$I(Z_i; Z') \;\le\; I(X_i; X') \;=\; R + U_i. \tag{23}$$

Since $X' \perp X_i \mid Y$, the representations satisfy $Z' \perp Z_i \mid Y$, and the data processing inequality applied to the chains $Z_i \to Y \to Z'$ and $Y \to X_i \to Z_i$ gives

$$I(Z_i; Z') \;\le\; I(Z_i; Y) \;\le\; I(X_i; Y) \;=\; R + U_i. \tag{24}$$

Step 3: Tightness under sufficient expressivity. As in the proof of Lemma 1, a map that is injective on the supports of X_i and X' attains the bound in Eq. (23), and the optimal parameters $\theta^\star$ that maximize $I(Z_i; Z')$ satisfy $I(Z_i; Z') = R + U_i$. Eq. (24) then holds with equality:

$$I(Z_i; Z') \;=\; I(Z_i; Y) \;=\; I(X_i; Y) \;=\; R + U_i. \tag{25}$$

The argument with Z'' in place of Z' is identical. □

Relation to CoMM. CoMM states the unimodal result as the special case of its multimodal lemma in which the augmentation is the projection t_i. Under Assumption 1, $t_i \in \mathcal{T}$ would require $I(X_i; Y) = I(X; Y)$, which by Eqs. (1) and (2) holds only when $S = 0$ and $U_j = 0$ for $j \neq i$. The proof above instead transfers Assumption 1 from X to X_i in Step 1 and does not require $t_i \in \mathcal{T}$.

PROOF OF LEMMA 3

Proof. Recall $\hat{Z} = P([Z_1; Z_2])$ and $Z_{\text{res}} = Z - \hat{Z}$, so $Z = \hat{Z} + Z_{\text{res}}$ pointwise.

Step 1: Equivalence under conditioning on $\hat{Z}$. Given $\hat{Z}$, the residual Z_{res} determines Z via $Z = \hat{Z} + Z_{\text{res}}$, and conversely Z determines Z_{res} via $Z_{\text{res}} = Z - \hat{Z}$. The two random variables therefore carry the same information given $\hat{Z}$, hence

$$I(Z_{\text{res}}; Y \mid \hat{Z}) \;=\; I(Z; Y \mid \hat{Z}). \tag{26}$$

Step 2: Chain rule decomposition. The chain rule for mutual information gives

$$I(Z, \hat{Z}; Y) \;=\; I(\hat{Z}; Y) + I(Z; Y \mid \hat{Z}). \tag{27}$$

Step 3: Bounds on $I(Z, \hat{Z}; Y)$. Both $Z = f_\theta(X)$ and $\hat{Z} = P([Z_1; Z_2]) = P([f_\theta(t_1(X)); f_\theta(t_2(X))])$ are deterministic functions of X, so the joint $(Z, \hat{Z})$ is a function of X. The data processing inequality therefore yields

$$I(Z, \hat{Z}; Y) \;\le\; I(X; Y). \tag{28}$$

The reverse direction follows from the chain rule, $I(Z, \hat{Z}; Y) = I(Z; Y) + I(\hat{Z}; Y \mid Z)$, and the non-negativity of conditional mutual information:

$$I(Z, \hat{Z}; Y) \;\ge\; I(Z; Y). \tag{29}$$

When $I(Z; Y) = I(X; Y)$, Eqs. (28)–(29) give

$$I(Z, \hat{Z}; Y) \;=\; I(Z; Y). \tag{30}$$

Step 4: Combine. Combining Eq. (29) with Eq. (27) and rearranging,

$$I(Z; Y \mid \hat{Z}) \;\ge\; I(Z; Y) - I(\hat{Z}; Y), \tag{31}$$

with equality under Eq. (30). By Eq. (26), the left side equals $I(Z_{\text{res}}; Y \mid \hat{Z})$, completing the proof. □

PROOF OF THEOREM 1

Proof. Apply Lemma 3 to the MSE-optimal projector $P^\star$ with corresponding prediction $\hat{Z}^\star$ and residual $Z^\star_{\text{res}} = Z - \hat{Z}^\star$:

$$I(Z^\star_{\text{res}}; Y \mid \hat{Z}^\star) \;\geq\; I(Z;Y) - I(\hat{Z}^\star; Y). \tag{32}$$

The PID decomposition in Eq. (1) and the definition of Δ give

$$I(Z;Y) \;=\; I(X;Y) - \Delta \;=\; R + S + U_1 + U_2 - \Delta. \tag{33}$$

Hypothesis 1 bounds the task-relevant content of the linear prediction:

$$I(\hat{Z}^\star; Y) \;\leq\; R + U_1 + U_2. \tag{34}$$

Substituting Eqs. (33)–(34) into Eq. (32),

$$I(Z^\star_{\text{res}}; Y \mid \hat{Z}^\star) \;\geq\; (R + S + U_1 + U_2 - \Delta) - (R + U_1 + U_2) \;=\; S - \Delta. \tag{35}$$

When $\Delta = 0$, Eq. (35) gives $I(Z^\star_{\text{res}}; Y \mid \hat{Z}^\star) \geq S$, so the residual provides at least S additional task-relevant information given the linear prediction. □

PROOF OF COROLLARY 1

Proof. Let $X' = t'(X)$ and $X'' = t''(X)$ for $t', t'' \in \mathcal{T}$, with corresponding fused representations $Z' = f_{\theta^\star}(X')$ and $Z'' = f_{\theta^\star}(X'')$, unimodal representations $Z'_i = f_{\theta^\star}(t_i(X'))$ and $Z''_i = f_{\theta^\star}(t_i(X''))$, and residuals

$$Z'_{\text{res}} \;=\; Z' - P^\star([Z'_1; Z'_2]), \qquad Z''_{\text{res}} \;=\; Z'' - P^\star([Z''_1; Z''_2]). \tag{36}$$

Step 1: Task-relevant information is preserved across views. By Assumption 1, $I(X';Y) = I(X'';Y) = I(X;Y)$, so the total task-relevant information $R + S + U_1 + U_2$ is preserved in both views. By Lemma 5, the fused representations satisfy $I(Z';Y) = I(Z'';Y) = I(X;Y)$ at the information-preserving optimum, which gives $\Delta = 0$ in Theorem 1 for each view.

Step 2: Conditional task-relevant information in the residuals. By Theorem 1 applied to each view,

$$I(Z'_{\text{res}}; Y \mid \hat{Z}'^\star) \;\geq\; S, \qquad I(Z''_{\text{res}}; Y \mid \hat{Z}''^\star) \;\geq\; S, \tag{37}$$

where $\hat{Z}'^\star$ and $\hat{Z}''^\star$ are the corresponding linear predictions. Given the linear prediction, each residual therefore provides at least S additional task-relevant information.

Step 3: InfoNCE rewards task-relevant information shared between the residuals. The InfoNCE estimator $\hat{I}_{\text{NCE}}(Z'_{\text{res}}, Z''_{\text{res}})$ is a lower bound on $I(Z'_{\text{res}}; Z''_{\text{res}})$, and maximizing this lower bound rewards information shared between the two residuals. Each residual is a deterministic function of its corresponding augmented view, and $X' \perp X'' \mid Y$ by Lemma 5. Consequently, $Z'_{\text{res}} \perp Z''_{\text{res}} \mid Y$, and the data processing inequality gives

$$I(Z'_{\text{res}}; Z''_{\text{res}}) \;\leq\; \min\left\{I(Z'_{\text{res}}; Y), I(Z''_{\text{res}}; Y)\right\}. \tag{38}$$

The information shared between the interaction residuals is therefore task-relevant under Assumption 1. By Step 2, each residual provides at least S additional task-relevant information given its linear prediction. Applying InfoNCE to the interaction residuals encourages the encoder to preserve shared task-relevant information in the component remaining after subtraction of the linear prediction, providing the motivation for synergy-focused supervision. □

E APPENDIX DISCUSSION OF THE AUGMENTATION ASSUMPTION

We discuss the interpretation and feasibility of Assumption 1, its role in SynCo, and the effect of an approximate version on the guarantee of Theorem 1.

Interpretation. Lemma 4 restates Assumption 1 as two conditional independences. The condition $X' \perp X \mid Y$ requires the augmented view to carry no information about X beyond the label, and the condition $Y \perp X \mid X'$ requires the view to retain all information that X carries about the label. Practical augmentation families satisfy both conditions only approximately. The assumption formalizes the goal of view design in contrastive learning, where augmentations remove nuisance information and preserve task-relevant information (Tian et al., 2020b), and extends the goal to multimodal augmentations that are jointly defined on (X_1, X_2).

Feasibility. CoMM examines the feasibility of the assumption by varying the augmentation strength on Trifeature and MM-IMDb (Dufumier et al., 2024, Appendix C.4). Downstream performance improves as stronger augmentations remove nuisance information and declines once the augmentations also remove task-relevant information. The interior optimum corresponds to the regime that Assumption 1 describes.

Role in SynCo. Assumption 1 underlies Lemmas 1 and 2, which characterize the optimum of $\mathcal{L}_{\text{MCL}}$ in Eq. (4). CoMM, InfMasking (Wen et al., 2025), and SynCo all train with $\mathcal{L}_{\text{MCL}}$, so the assumption applies equally to all three methods and does not differentiate SynCo from them. Lemma 3 and the bound of Theorem 1 do not require Assumption 1; the assumption enters only through $\Delta = 0$ in Theorem 1. The premise specific to SynCo is Hypothesis 1, which concerns the trained encoder and the linear projector rather than the augmentation family and is assessed empirically in Table 2.

Approximate label preservation. When the augmentations preserve task-relevant information only approximately, Δ can be positive, and Theorem 1 still gives $I(Z^\star_{\text{res}}; Y \mid \hat{Z}^\star) \geq S - \Delta$.

F Appendix Broader Impact

SynCo is a method for self-supervised multimodal representation learning that improves the capture of synergistic information across modalities. As foundational research on contrastive learning objectives, the work is not tied to a specific application or deployment, but its downstream applications span both beneficial and potentially sensitive domains.

Potential positive impacts. By improving the quality of representations learned from unlabeled multimodal data, SynCo can reduce the labeling burden in domains where annotations are expensive or require expert knowledge. The benchmarks evaluated in this paper highlight several such domains: clinical decision support from physiological time-series and tabular records (MIMIC), human activity recognition from wearable sensors (DARai), and robotic manipulation from vision and force-torque feedback (Vision&Touch). Improved synergy capture is particularly valuable in these settings, where the most informative cues often emerge only from the joint observation of multiple modalities. SynCo also introduces negligible computational overhead relative to existing contrastive frameworks (Appendix C), lowering the energy cost and hardware barrier to applying multimodal self-supervised learning compared to more compute-intensive alternatives.

Potential negative impacts. As with any general-purpose representation learning method, SynCo could in principle be applied in settings with negative consequences, such as surveillance applications. The method is presented as a general-purpose training objective rather than a deployable system, and we do not release any pretrained model that constitutes a direct path to such applications.

Fairness and reliability. Models trained with SynCo inherit the properties of their training data and backbone encoders. In high-stakes domains, deployment should follow established validation and evaluation protocols, which are independent of the training objective.

G Appendix Dataset Details

G.1 Trifeature

The Trifeature benchmark (Hermann & Lampinen, 2020) was originally introduced for studying how convolutional networks represent visual factors of variation. Each image contains a single shape

rendered with a chosen texture and a chosen color, and shape, texture, and color each take one of ten values, giving $1{,}000$ unique (shape, texture, color) combinations. We assign 800 combinations to training and reserve the remaining 200 for evaluation. Every training combination is rendered three times, with the shape and texture orientations sampled independently from $[-45°, 45°]$. The shape occupies a 128×128 region placed at a uniformly random location inside a 224×224 image canvas so that the region lies entirely within the canvas, and the texture and color are then applied. We form bimodal examples by pairing two such images, producing $10{,}000$ training pairs and $4{,}096$ test pairs from the same distribution. Section 5.1 describes how each pair is labeled to obtain the redundancy, uniqueness, and synergy probes.

G.2 MULTIBENCH

We use five datasets from the MultiBench benchmark (Liang et al., 2021) and follow the preprocessing released with the benchmark in every case. All datasets are anonymized at source and contain no personally identifiable information.

MIMIC. The MIMIC dataset (Johnson et al., 2016) comprises de-identified critical-care records gathered between 2001 and 2012, totaling $53{,}423$ admissions from $38{,}597$ distinct patients. We use the two modalities released by MultiBench: a time-series modality of hourly 12-dimensional physiological measurements taken over the first 24 hours of an admission, and a static modality of 5-dimensional patient demographics such as age and gender. Consistent with (Liang et al., 2023b), the target is a binary indicator of whether an admission falls under ICD-9 code group 7 (460-519), corresponding to diseases of the respiratory system.

MOSI. MOSI (Zadeh et al., 2016) is a sentiment analysis benchmark built from $2{,}199$ short YouTube monologue clips, each accompanied by video frames, audio, and a transcript of the spoken content. The original labels are continuous sentiment scores in $[-3, 3]$, which we binarize into positive and negative classes consistent with (Liang et al., 2023b). Our bimodal experiments use the visual and textual modalities.

UR-FUNNY. UR-FUNNY (Hasan et al., 2019) is a benchmark for humor detection built from $1{,}866$ TED talks, providing $16{,}514$ short video segments with aligned audio and transcripts, each labeled as humorous or not. Our bimodal experiments use the visual and textual modalities.

MUSTARD. MUSTARD (Castro et al., 2019) is a sarcasm detection benchmark comprising utterances from popular television shows, including *Friends*, *The Big Bang Theory*, and *The Golden Girls*. We use the balanced split of 690 utterances, with every utterance providing synchronized video, audio, and transcript along with a binary sarcasm label. Our bimodal experiments use the visual and textual modalities.

Vision&Touch. Vision&Touch (Lee et al., 2020) records 150 trajectories of $1{,}000$ time steps each, captured from a 7-DoF Franka Panda robot performing a peg-insertion task. Every step provides an RGB image, a depth map, force readings, and the end-effector position and velocity. Consistent with the regression setting in MultiBench (Liang et al., 2021), we predict the next-step end-effector position under mean-squared error. Our bimodal experiments use the visual and proprioceptive modalities.

G.3 DARAI

DARai (Kaviani et al., 2025) is a multimodal benchmark for human activity recognition, providing over 200 hours of synchronized recordings from 50 participants in 10 indoor environments. The full release contains 20 sensor streams, including multiple camera views, depth and radar sensors, wearable inertial measurement units (IMU), forearm electromyography (EMG), instrumented insoles, biomonitors, and a gaze tracker. Annotations are organized in a three-level hierarchy spanning high-level activities, lower-level actions, and fine-grained procedures. We use DARai to assess whether the gains of SynCo extend beyond the vision-language regime that dominates the other benchmarks. The wearable subset operates on physical signals with markedly different temporal dynamics and information content from natural images and text.

Each wearable modality captures a distinct physical aspect of human activity. The wearable IMU records three-axis accelerometer, gyroscope, and magnetometer signals at 12 Hz and reflects the kinematics of the upper body, including limb orientation, acceleration, and rotation. The forearm EMG records muscle activity at 4 kHz and carries information about exerted effort that is largely invisible to a purely kinematic sensor, since the same motion can arise from very different muscular loads. The instrumented insoles record eight-point pressure at 500 Hz and capture weight transfer, balance, and ground-contact patterns that the upper-body sensors cannot observe. The wearable subset thus pairs signals that differ in physical origin, body location, and temporal scale.

To assess whether the gains of SynCo depend on the choice of modality pairing, we evaluate three bimodal configurations: IMU + Insole, IMU + EMG, and EMG + Insole. The IMU + Insole pairing combines arm-mounted kinematics with foot-mounted pressure. Activities requiring coordinated upper- and lower-body motion, such as climbing stairs or carrying an object, can only be inferred by integrating both signals. The IMU + EMG pairing relates gross kinematics to muscular effort and exposes activities where the same motion arises from different effort profiles, such as lifting a heavy versus a light object. The EMG + Insole pairing combines forearm and foot signals from physically separated body regions, requiring the model to relate cues across distant sensors rather than rely on a single body segment. The three configurations share the same preprocessing and backbone networks, as reported in Table 6.

G.4 MM-IMDB

MM-IMDb (Arevalo et al., 2017) is a multilabel benchmark for movie genre classification, with 25,959 films described by a poster image and a plot summary. The genre annotations span 23 categories, and a single film often belongs to several of them, which makes the task multilabel rather than multiclass. We use the poster and plot modalities. Although MM-IMDb is included in MultiBench (Liang et al., 2021), we report it separately because we train directly on the raw modalities rather than on the preprocessed feature vectors that MultiBench provides.