InfMasking: Unleashing Synergistic Information by Contrastive Multimodal Interactions

Liangjian Wen[1,2,8*], Qun Dai[1*], Jianzhuang Liu[3], Jiangtao Zheng[1], Yong Dai[4‡], Dongkai Wang[1†], Zhao Kang[5], Jun Wang[1], Zenglin Xu[6,7], Jiang Duan[1†]

[1]School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics
[2]Engineering Research Center of Intelligent Finance, Ministry of Education,Southwestern University of Finance and Economics
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[4]X-Humanoid
[5]University of Electronic Science and Technology of China
[6]Shanghai Academy of AI for Science
[7]Artificial Intelligence Innovation and Incubation Institute, Fudan University
[8]Artificial Intelligence and Digital Finance Key Laboratory of Sichuan Province
{wlj6816,daiqun1124}@gmail.com

Abstract

In multimodal representation learning, synergistic interactions between modalities not only provide complementary information but also create unique outcomes through specific interaction patterns that no single modality could achieve alone. Existing methods may struggle to effectively capture the full spectrum of synergistic information, leading to suboptimal performance in tasks where such interactions are critical. This is particularly problematic because synergistic information constitutes the fundamental value proposition of multimodal representation. To address this challenge, we introduce InfMasking, a contrastive synergistic information extraction method designed to enhance synergistic information through an **Inf**inite **Masking** strategy. InfMasking stochastically occludes most features from each modality during fusion, preserving only partial information to create representations with varied synergistic patterns. Unmasked fused representations are then aligned with masked ones through mutual information maximization to encode comprehensive synergistic information. This infinite masking strategy enables capturing richer interactions by exposing the model to diverse partial modality combinations during training. As computing mutual information estimates with infinite masking is computationally prohibitive, we derive an InfMasking loss to approximate this calculation. Through controlled experiments, we demonstrate that InfMasking effectively enhances synergistic information between modalities. In evaluations on large-scale real-world datasets, InfMasking achieves state-of-the-art performance across seven benchmarks. Code is released at `https://github.com/brightest66/InfMasking`.

*Equal contribution.
†Corresponding author.
‡Project leader.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Multimodal contrastive learning has revolutionized representation learning by enabling the integration of diverse data modalities, such as text, images, and audio, into a unified latent space. This paradigm leverages contrastive loss [37, 9] to align features from different modalities as pioneered by foundational works like CLIP [38] and ALIGN [23] in vision-language tasks. These models demonstrate the power of aligning multimodal features to capture shared patterns across data sources, enabling versatile downstream applications. However, current approaches often rely on the restrictive *multiview redundancy assumption* [40, 43, 46], which posits that a modality is (approximately) sufficient for the prediction of downstream tasks and contains the same task-relevant information. This assumption derives from multi-view learning and is limited in real-world multimodal settings because many multimodal tasks involve minimal shared information.

The shortcomings of this redundancy-centric perspective become increasingly apparent when we examine the multifaceted nature of multimodal interactions. As illustrated in [6, 12], these interactions can be classified into three fundamental categories: redundancy, uniqueness, and synergy. Redundancy refers to scenarios where a modality can independently perform the task due to overlapping, shared information. Uniqueness describes cases where only one modality possesses all the requisite information for task completion. Synergy, arguably the most significant yet elusive of the three, occurs when modalities provide complementary information that must be combined to achieve the desired outcome. These interaction types are not static; their predominance varies depending on the specific task, adding a layer of complexity to multimodal learning. For instance, a task might rely heavily on redundant information in one context, while another demands the synergistic integration of modalities to succeed. A compelling example of this is hateful meme detection [26], where synergy emerges when seemingly neutral modalities (such as an innocuous image and benign text) combine to create harmful content that neither conveys on its own. This highlights the critical importance of synergistic integration, as models must fuse information from different modalities cues to uncover implicit biases or offensive implications that are only apparent in their joint context, enabling more effective identification and mitigation of such content in real-world applications. Consequently, task-agnostic multimodal representations must necessarily encompass the full spectrum of multimodal interactions that transcend mere informational redundancy.

Recently, FactorCL [29] explicitly decomposes shared and unique representations, enabling the estimation of redundancy and unique interactions beyond multi-view redundancy. However, its factorized approximation of multimodal interactions is prone to cumulative errors and fails to capture synergistic information effectively. In contrast, [12] integrates features from all modalities to derive a common representation and subsequently maximizes the mutual information between this representation and its augmented variants, as well as between the common representation and its corresponding unimodal features. Although this approach facilitates the capture of redundant, unique, and synergistic information across modalities, it primarily emphasizes enhanced redundant and unique interactions. Synergistic information is captured by maximizing the mutual information between the common representation and its augmented variants. Such handling may prove inadequate for tasks that heavily rely on complex inter-modal synergy. The complexity of synergistic interactions lies not merely in modalities providing complementary information but in how these modalities combine through specific interaction patterns to produce outcomes unattainable by any single modality alone. Such interactions may involve nonlinear dependencies or context-dependent dynamics. Hence, effectively capturing the full spectrum of synergistic information remains a significant challenge.

To address this challenge, we introduce a contrastive synergistic information extraction method via infinite masking (denoted as InfMasking). Specifically, InfMasking stochastically occludes a substantial proportion of features from each modality during the fusion process. This masking preserves only partial information, creating fused representations with varied synergistic patterns. Subsequently, unmasked fused representations are aligned with these masked ones via mutual information maximization to encode comprehensive synergistic information. The infinite masking strategy enables InfMasking to capture richer synergistic interactions by exposing the model to diverse combinations of partial modality information during training. However, the computation of mutual information estimates with infinite masking is computationally prohibitive. To address this issue, we derive an InfMasking loss to approximate the calculation of this loss function. Empirically, InfMasking effectively captures synergistic information between modalities in controlled environments with known interaction types. When tested on real-world datasets across diverse domains (healthcare, robotics) and data types (image, text, audio), InfMasking achieves state-of-the-art results on seven multimodal tasks involving two or three modalities.

2 Preliminary: Contrastive Multimodal Interactions

Consider $X_1, X_2, \ldots, X_n$ as random variables, each representing a distinct modal data (e.g., image, text, audio, or tabular data), alongside a downstream task Y. The objective is to derive an effective multimodal latent variable $Z_\theta = f_\theta(X)$, where $X = (X_1, \ldots, X_n)$ and θ detnotes the parameter of multimodal encoder. Multimodal interactions refer to the process of extracting and integrating information from multiple data sources, such as text, image, audio, or tabular data, to form a cohesive representation for downstream tasks. These interactions can be categorized into three types: redundancy (R), where information is shared across modalities; uniqueness (U), where information is specific to a single modality; and synergy (S), where complementary information emerges only when modalities are combined.

To understand multimodal interactions, we can leverage Partial Information Decomposition (PID) [49, 6], a theoretical framework that dissects the mutual information between multiple variables. For analytical tractability, we focus on the case of $n = 2$. Consider two modalities X_1 and X_2 and a task Y. The mutual information $I(X_1, X_2; Y)$ quantifies the total task-relevant information provided by both modalities jointly. According to PID, this can be decomposed as: $I(X_1, X_2; Y) = R + S + U_1 + U_2$, where R represents redundant information, common to both X_1 and X_2, S represents synergistic information, emerging only from the combination of X_1 and X_2, and U_1 and U_2 represent unique information specific to X_1 and X_2, respectively. This decomposition is supported by consistency equations derived from the chain rule of mutual information:

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

Effectively capturing these interactions is fundamental to multimodal learning, as they collectively enhance a model's ability to generalize across diverse tasks. Therefore, an effective multimodal representation Z_θ must capture information relevant to a downstream task Y, preserving the mutual information such that $I(Z; Y) = I(X; Y)$. In self-supervised learning, however, Y remains unspecified during the representation learning phase, presenting a unique challenge. To address this, [12] extends the contrastive learning principle of multiview redundancy to multimodal contexts:

Assumption 1 (*Minimal label-preserving multimodal augmentations*) *A set $\mathcal{T}^*$ of multimodal transformations exists, such that for any $t \in \mathcal{T}^*$ and $X' = t(X)$, the mutual information $I(X; X') = I(X; Y)$ holds, preserving the information with label Y.*

Assumption 1 is justified within the framework of multimodal representation learning. It enables a broader range of augmentations, as it is not limited to the set $\mathcal{T}_c^\star = \{t(X) = (t_1(X_1), t_2(X_2)), \ldots, t_n(X_n)\}$. We define $Z'_\theta = f_\theta(X')$, where $X' = t(X)$ with $t \in \mathcal{T}$ representing multimodal augmentation. Considering the data processing inequalities for the Markov chains $X \to X' \to Z'_\theta$ and $Z'_\theta \to X \to Z_\theta$, we can establish the following mutual information bounds:$I(Z_\theta; Z'_\theta) \leq I(X; Z'_\theta) \leq I(X; X')$. According to these inequalities and Assumption 1, we can prove the following lemmas:

Lemma 1 *When optimizing the function f_θ to maximize mutual information $I(Z_\theta; Z'_\theta)$, and under the assumption that the network f_θ possesses sufficient expressivity, we observe that in the optimal parameter configuration: $I(Z_{\theta^\star}, Z'_{\theta^\star}) = I(X, X') = I(X, Y)$.*

For Z_θ to serve as an effective representation of X, it must adequately preserve and encode all task-relevant information inherent in X. We note that $I(X; Y) = I(X_1, X_2; Y) = R + S + U_1 + U_2$. Based on Lemma 1, we can learn common multimodal representations Z_θ and quantify all multimodal interactions beyond redundancy by maximizing the mutual information $I(Z_\theta; Z'_\theta)$.

Lemma 2 *Suppose f_{θ^*} is optimal, meaning it maximizes $I(Z_{\theta^\star}, Z'_{\theta^\star})$. Then, the equality $I(Z_{\theta^\star}, Y) = I(X', Y)$ holds. For the special case where $T = \{t_i\}$ such that $X' = t_i(X) = X_i$ and $Z'_{\theta^\star} = f_{\theta^\star}(X) = Z_i$ for $i \in \{1, 2\}$, the following holds:$I(Z_i; Y) = I(X_i; Y) = R + U_i$.*

For unimodal representations Z_i where $i \in \{1, 2\}$ to effectively represent X_i, each representation must encode the information $I(X_i; Y) = R + U_i$ contained in the corresponding modality. According to Lemma 2, we can learn optimal unimodal representations Z_i and quantify both redundant and unique multimodal interactions by maximizing the mutual information $I(Z_i; Z'_\theta)$ and $I(Z_i; Z_\theta)$.

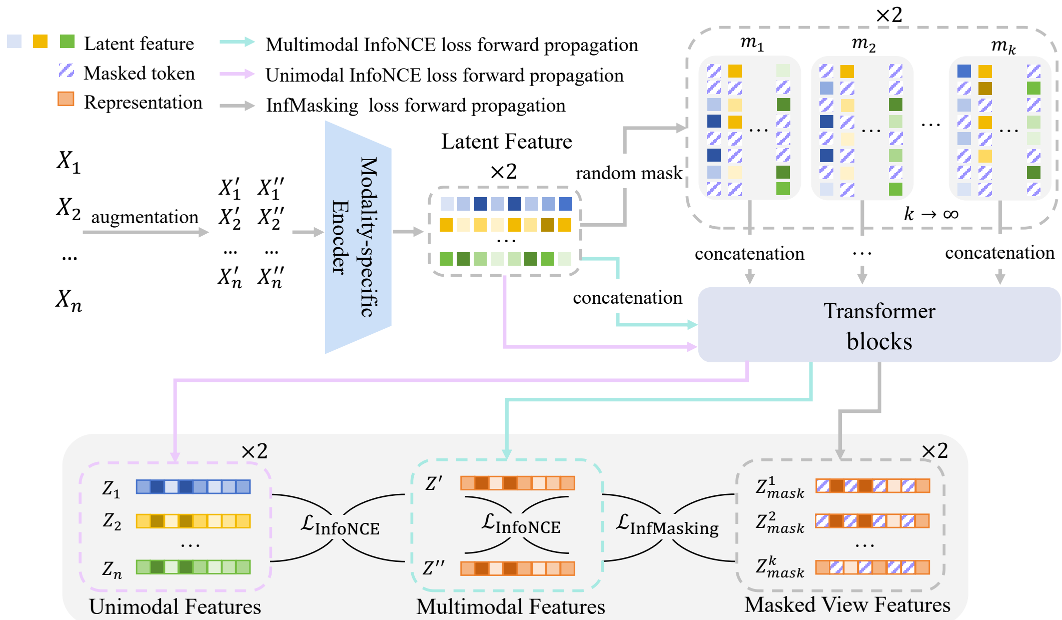


Figure 1: The overall pipeline of InfMasking. Given n modalities $X = (X_1, X_2, \dots, X_n)$, we augment them to obtain X' and X'', which are then encoded independently by modality-specific encoders to extract latent features. These features are processed in three ways: (1) All modality features are concatenated and input into a Transformer block, yielding fused features Z' and Z''; (2) Each modality feature is individually input into a Transformer block, producing unimodal features $Z_1, Z_2, \dots, Z_n$; (3) Features of each modality are randomly masked, concatenated, and input into a Transformer block, repeated k times to obtain $Z^1_{\text{mask}}, Z^2_{\text{mask}}, \dots, Z^k_{\text{mask}}$.

3 Unleashing Synergistic Information through Infinite Masking

3.1 The General Framework

The proposed framework, termed InfMasking, is a multimodal contrastive interaction method designed to enhance synergistic information across modalities by leveraging infinite masked views. The overall pipeline of InfMasking is illustrated in Fig. 1 and consists of two primary stages: modality-specific latent feature encoding and multimodal feature fusion via a Transformer. Given an input set of n modalities $X = (X_1, X_2, \dots, X_n)$, we obtain X' and X'' through an augmentation process. Subsequently, X' and X'' are processed by modality-specific encoders, where each modality is encoded independently to extract latent features. As shown in Fig. 1, these modality features are then processed in parallel through three distinct ways: (1) All modality features are concatenated and input into a Transformer block, yielding fused features Z' and Z''; (2) Each modality feature is individually input into a Transformer block, producing unimodal features $Z_1', Z_2', \dots, Z_n'$ and $Z_1'', Z_2'', \dots, Z_n''$; (3) Features of each modality are randomly masked, then concatenated and input into a Transformer block to obtain fused features. This process is executed k times, resulting in $Z'^1_{\text{mask}}, Z'^2_{\text{mask}}, \dots, Z'^k_{\text{mask}}$ and $Z''^1_{\text{mask}}, Z''^2_{\text{mask}}, \dots, Z''^k_{\text{mask}}$.

Based on Lemma 1 and Lemma 2, [12] proposes a contrastive multimodal (CoMM) learning approach to learn task-agnostic multimodal representations by modeling multimodal interactions, including redundancy, uniqueness, and synergy. CoMM estimates the mutual information using the InfoNCE loss: $\hat{I}_{\text{NCE}}(Z, Z') = \mathbb{E}_{z, z'_{\text{pos}} \sim p(Z, Z')} \left[\log \frac{\exp(z^T z'_{\text{pos}}/\tau)}{\exp(z^T z'_{\text{pos}}/\tau) + \sum_{z'_{\text{neg}}} \exp(z^T, z'_{\text{neg}}/\tau)} \right]$, where τ is a temperature parameter. Hence, its training objective is formulated as follows:

$$\mathcal{L}_{\text{CoMM}} = - \underbrace{\hat{I}_{\text{NCE}}(Z', Z'')}_{\approx R + S + \sum_{i=1}^{n} U_i} - \sum_{i=1}^{n} \frac{1}{2} \underbrace{\left(\hat{I}_{\text{NCE}}(Z_i, Z') + \hat{I}_{\text{NCE}}(Z_i, Z'') \right)}_{\approx R + U_i} . \tag{1}$$

While the first term theoretically quantifies redundancy, synergy, and the unique information across modalities, empirical evidence indicates that its practical performance exhibits notable limitations. Hence, the second term constitutes the fundamental component of CoMM, specifically designed to strengthen both unique and redundant interaction patterns. However, enhancing synergistic interactions remains a substantial challenge in this framework.

Synergy is a complex interaction that arises when different modalities provide complementary information, necessitating their integration for effective task performance. We aim to learn a multimodal representation that captures all three types of interactions, with a particular emphasis on enhancing synergistic information. To achieve this, we introduce InfMasking , a novel approach that leverages infinite masking to enhance the modeling of multimodal interactions. Our training objective is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{Total loss}} = &- \underbrace{\hat{I}_{\text{NCE}}(Z', Z'')}_{\approx R+S+\sum_{i=1}^{n} U_i} - \sum_{i=1}^{n} \frac{1}{2} \underbrace{\left(\hat{I}_{\text{NCE}}(Z_i, Z') + \hat{I}_{\text{NCE}}(Z_i, Z'')\right)}_{\approx R+U_i} \\ &\underbrace{- \mathbb{E}_{\text{mask}}\left[\hat{I}_{\text{NCE}}(Z'_{\text{mask}}, Z') + \hat{I}_{\text{NCE}}(Z''_{\text{mask}}, Z'')\right]}_{\mathcal{L}_{\text{InfMasking}}}, \end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{InfMasking}}$ represents our novel masking-based regularization term designed to specifically enhance synergistic interactions, as detailed in Section 3.2.

3.2 Contrastive Synergistic Information via Infinite Masking

In multimodal learning, capturing synergistic information—where different modalities provide complementary insights—is essential for tasks requiring integrated understanding. We propose a contrastive synergistic Information method via infinte masking to enhance synergistic interactions. It's core idea is to randomly mask a significant portion of the features from each modality during the fusion process. As shown in Fig. 1, we fuse all masked features from different modalities to obtain a fused representation Z'_{mask} via a Transformer. Each time features from each modality are randomly masked, only partial information from each modality is retained. Consequently, after each masking operation, Z'_{mask} contains distinct synergistic information. Then, by aligning Z'_{mask} with its unmasked counterparts Z' through maximizing their mutual information, Z' are encouraged to capture distinct synergistic information. This process is repeated for K times of masking, we can obtain the final training loss: $\frac{1}{K}\sum_{k=1}^{K} \hat{I}_{\text{NCE}}(Z'^{\,k}_{\text{mask}}, Z') + \hat{I}_{\text{NCE}}(Z''^{\,k}_{\text{mask}}, Z'')$. To enable the model to learn diverse forms of synergistic information, we allow K to approach infinity through infinite masking, ultimately obtaining the masking loss $\mathcal{L}_{\text{InfMasking}}$ as follows:

$$\begin{aligned}\mathcal{L}_{\text{InfMasking}} &= - \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \hat{I}_{\text{NCE}}(Z'^{\,k}_{\text{mask}}, Z') + \hat{I}_{\text{NCE}}(Z''^{\,k}_{\text{mask}}, Z'') \\ &= -\mathbb{E}_{\text{mask}}\left[\hat{I}_{\text{NCE}}(Z'_{\text{mask}}, Z') + \hat{I}_{\text{NCE}}(Z''_{\text{mask}}, Z'')\right]. \end{aligned} \tag{3}$$

This infinite masking strategy enables InfMasking to capture richer synergistic interactions by exposing the model to diverse combinations of partial modality information during training. However, the estimation of $\mathbb{E}_{\text{mask}}\left[\hat{I}_{\text{NCE}}(Z'_{\text{mask}}, Z')\right]$ and $\mathbb{E}_{\text{mask}}\left[\hat{I}_{\text{NCE}}(Z''_{\text{mask}}, Z'')\right]$ is computationally expensive via random mask samples.

To address this issue, we derive a lower bound for $\mathbb{E}_{\text{mask}}\left[\hat{I}_{\text{NCE}}(Z'_{\text{mask}}, Z')\right]$ to optimize the InfMasking loss function Eq. (3). The detailed derivation is as follows:

$$\mathbb{E}_{\text{mask}}[\hat{I}_{\text{NCE}}(Z'_{\text{mask}}, Z')] = \mathbb{E}_{\text{mask}}[\mathbb{E}_{z'\sim p(Z')}\left[\log \frac{\exp(z'^T z'_{\text{mask}}/\tau)}{\exp(z'^T z'_{\text{mask}}/\tau) + \sum_{z'_{\text{neg}}} \exp(z'^T z'_{neg}/\tau)}\right]] \quad (4)$$

$$= \mathbb{E}_{z'\sim p(Z')}[\mathbb{E}_{\text{mask}}\left[(z'^T z'_{\text{mask}}/\tau) - \log[\exp(z'^T z'_{\text{mask}}/\tau) + \sum_{z'_{\text{neg}}} \exp(z'^T z'_{neg}/\tau)]\right]] \quad (5)$$

$$\geq \mathbb{E}_{z'\sim p(Z')}\left[z'^{\,T}\mathbb{E}_{\text{mask}}[z'_{\text{mask}}]/\tau - \log \mathbb{E}_{\text{mask}}[\exp(z'^T z'_{\text{mask}}/\tau) + \sum_{z'_{\text{neg}}} \exp(z'^T z'_{neg}/\tau)]\right] \quad (6)$$

The inequality Eq. (6) merges from the application of Jensen inequality on concave functions i.e., $\mathbb{E}_x \log(X) \leq \log \mathbb{E}_x[X]$. z'_{mask} denotes the integrated representation derived from the fusion of all masked features across diverse modalities via the Transformer architecture.

Inspired by [7], we posit that z'_{mask} follows a Gaussian distribution, formally expressed as $z'_{\text{mask}} \sim \mathcal{N}(\boldsymbol{\mu}_{z'_{\text{mask}}}, \boldsymbol{\Sigma}_{z'_{\text{mask}}})$, where $\boldsymbol{\mu}_{z'_{\text{mask}}}$ and $\boldsymbol{\Sigma}_{z'_{\text{mask}}}$ denote the mean vector and covariance matrix of z'_{mask}, respectively. This assumption is well-founded for two principal reasons. First, the masked embeddings tend to cluster around a central value in the embedding space, as they all inherently reflect aspects of the query's semantic nature. Second, the variance observed across feature dimensions can be interpreted as a representation of semantic differentiation in the ambient space, which aligns with established principles in distributional semantics. Under this assumption, we can derive:

Lemma 3 *Let $\boldsymbol{\mu}_{z'_{mask}}$ and $\boldsymbol{\Sigma}_{z'_{mask}}$ be the Gaussian mean vector and covariance matrix of z'_{mask}, respectively. The lower bound of $\mathbb{E}_{mask}\left[\hat{I}_{NCE}(Z'_{mask}, Z')\right]$ can be obtained as follows:* $\mathbb{E}_{mask}[\hat{I}_{NCE}(Z'_{mask}, Z')]$

$$\geq \mathbb{E}_{z'\sim p(Z')}\left[z'^T \boldsymbol{\mu}_{z'_{mask}}/\tau - \log[\exp(z'^T \boldsymbol{\mu}_{z'_{mask}}/\tau + \frac{z'^T \boldsymbol{\Sigma}_{z'_{mask}} z'}{2\tau^2}) + \sum_{z'_{neg}} \exp(z'^T z'_{neg}/\tau)]\right] \quad (7)$$

This allows us to approximate the mutual information between the masked and unmasked representations without requiring exhaustive sampling of all possible masks. A detailed proof is given in Appendix G.

4 Experiments

We perform experiments on both synthetic benchmarks and multiple large-scale real-world datasets to verify the effectiveness of InfMasking in learning representations from diverse modalities. To evaluate InfMasking's capacity to capture three essential aspects of multimodal interactions (*i.e.*, uniqueness, redundancy, and synergy), we generate synthetic data in a controlled environment based on the Trifeature dataset [22]. Furthermore, we assess the generalizability of InfMasking on several widely used multimodal benchmark datasets involving diverse modality combinations in real-world scenarios. These tasks span various domains (*e.g.*, healthcare, robotics, *etc.*) allowing for a thorough assessment of the model's representation capabilities across diverse modalities. Detailed experimental settings are provided in Appendix A. For evaluation, we use linear probing, *i.e.*, freezing the pre-trained feature extractor and training a linear classifier (or regressor, depending on the task) on top of the learned representations. The downstream task performance of the linear model serves as an indicator of the quality of the learned multimodal representations.

4.1 Synthetic Experiments on Trifeature Datasets

Following the experimental design of the Trifeature dataset in CoMM [12], we conduct controlled experiments on a synthetic dataset derived from Trifeature. We assess the model's capacity to learn uniqueness, redundancy and synergy through two separate experiments. In terms of uniqueness and redundancy, we define shapes as redundant features and textures as uniqueness features. And the task involves two subtasks: (1) identifying the shared shape between two images (redundancy) and (2) predicting the texture of the first image (or second image) (uniqueness). The random-guessing baselines in both cases corresponds to 10%. As for synergy, we artificially introduce a strong correlation between textures and colors by defining a mapping $\mathcal{M}$ in the training set (*e.g.*, blue maps to triangles, stripes to red), resulting in a 50% baseline for random guessing. The model is trained on image pairs that follow this mapping. The task is to determine whether a given image pair satisfies the mapping $Y = \mathbb{1}(\mathrm{texture}(X_1), \mathrm{color}(X_2) \in \mathcal{M})$, thereby evaluating the model's ability to capture synergistic interactions across modalities.

Table 1: Linear probing accuracy (in %) of redundancy (shape), uniqueness (texture) and synergy (color and texture) on Trifeature dataset. ♣ denotes results are from [12].

Model	*redundancy*↑	*uniqueness*↑	*synergy*↑
Cross♣ [38]	**100.0**	11.6	50.0
Cross+Self♣ [51]	99.7	86.9	50.0
FactorCL♣ [29]	99.8	62.5	46.5
MAE [19]	$99.8_{\pm 0.11}$	$82.4_{\pm 3.09}$	$50.1_{\pm 0.24}$
CoMM [12]	$99.9_{\pm 0.06}$	$86.8_{\pm 2.99}$	$71.4_{\pm 3.47}$
InfMasking (ours)	$99.9_{\pm 0.09}$	**90.6**$_{\pm 2.31}$	**77.0**$_{\pm 4.22}$

Experimental results are illustrated in Tab. 1. Cross-modality constraints based on the InfoNCE loss [38] ("Cross" model) achieve the best performance at capturing redundant information but struggle with uniqueness and synergy. FactorCL [29], self-supervised constraints on each encoder ("Cross + Self" [51]) and MAE [19] (implementation details are provided in Appendix D.1) improve on uniqueness but remain limited in modeling synergy. CoMM [12] performs well across all three interactions. However, it still has considerable room for improvement, particularly in capturing synergistic information. In comparison, InfMasking achieves superior performance in capturing both redundancy and synergy, outperforming CoMM by 3.8% and 5.6%, respectively.

4.2 Experiments on Real-world Datasets

We further evaluate the performance of our model on several real-world multimodal datasets provided by Multibench [30]. These datasets span diverse modality combinations and task types, providing a comprehensive benchmark to assess the model's ability to learn effective multimodal representations. Further dataset details are provided in Appendix B.

4.2.1 Experiments with 2 Modalities on Multibench

Following the data preprocessing procedure of previous work [29, 12], we conduct our experiments using the same encoders, modality configurations and train models based on encoded inputs with diverse modalities. We consider "Cross", "Cross+Self", FactorCL and CoMM as baselines for comparison. As presented in Tab. 2, InfMasking consistently achieves the best performance across all

Table 2: Linear probing MSE($\times 10^{-4}$) for regression task and top-1 accuracy (in %) for classification tasks on Multibench. ♣ denotes results are from [29]. * denotes average is only selected from the results of classification tasks.

Model	*Regression*	*Classification*				
	V&T EE↓	*MIMIC*↑	*MOSI*↑	*UR-FUNNY*↑	*MUSTARD*↑	**Average*** ↑
Cross♣ [38]	$33.09_{\pm 3.67}$	$66.7_{\pm 0.1}$	$47.8_{\pm 1.8}$	$50.1_{\pm 1.9}$	$53.5_{\pm 2.9}$	54.52
Cross+Self♣ [51]	$7.56_{\pm 0.31}$	$65.49_{\pm 0.0}$	$49.0_{\pm 1.1}$	$59.9_{\pm 0.9}$	$53.9_{\pm 4.0}$	57.07
FactorCL♣ [29]	$10.82_{\pm 0.56}$	$67.3_{\pm 0.0}$	$51.2_{\pm 1.6}$	$60.5_{\pm 0.8}$	$55.80_{\pm 0.9}$	58.7
CoMM [12]	$7.96_{\pm 2.13}$	$66.4_{\pm 0.41}$	$63.7_{\pm 2.5}$	$63.3_{\pm 0.51}$	$64.4_{\pm 1.1}$	64.45
InfMasking (ours)	**4.23**$_{\pm 0.37}$	**68.1**$_{\pm 0.42}$	**69.0**$_{\pm 1.2}$	**64.3**$_{\pm 0.9}$	**66.8**$_{\pm 2.5}$	**67.05**

benchmark datasets. In the binary classification tasks, InfMasking outperforms CoMM—the strongest baseline—by 1.7%, 5.3%, 1.0%, and 2.4% on the MIMIC, MOSI, UR-FUNNY, and MUSTARD datasets, respectively. For regression tasks, InfMasking also delivers superior results, achieving a lead of 3×10^{-4} in terms of MSE compared to the second-best model. These experimental results demonstrate the effectiveness of InfMasking in capturing bimodal interactions. Furthermore, its consistently strong performance across diverse datasets highlights the generalizability and robustness of InfMasking in real-world bimodal scenarios.

4.2.2 Experiments with 3 Modalities on Multibench

We evaluate the generalizability of InfMasking in learning multimodal representations beyond two modalities. Specifically, we conduct experiments on two datasets: Vision&Touch (for the contact prediction task, with visual, force, and proprioceptive modalities) and UR-FUNNY (with visual, text, and audio modalities). CMC [41] and CoMM are selected as baselines for comparison in the three-modality setting.

Table 3: Linear probing top-1 accuracy (in %) for classification tasks on Vision&Touch and UR-FUNNY. ♣ denotes results are from [12].

Model	*#Mod.*	*V&T CP*↑	*UR-FUNNY*↑
Cross	2	$86.3_{\pm 0.25}$	50.1♣
Cross+Self	2	$87.6_{\pm 0.26}$	59.9♣
CoMM	2	$87.0_{\pm 1.77}$	$63.3_{\pm 0.51}$
InfMasking (ours)	2	$88.5_{\pm 0.33}$	$64.3_{\pm 0.9}$
CMC♣ [41]	3	94.1	59.2
CoMM	3	$\mathbf{94.1}_{\pm 0.17}$	$64.8_{\pm 1.13}$
InfMasking (ours)	3	$94.1_{\pm 0.09}$	$\mathbf{65.6}_{\pm 1.15}$

The results are summarized in Tab. 3. To more intuitively assess the information gain introduced by incorporating a third modality, we additionally report results from bi-modal training scenarios using CoMM, "Cross" and "Cross + Self". Specifically, we train these baselines on (1) the image and proprioceptive modalities of the Vision&Touch dataset, and (2) the image and text modalities of the UR-FUNNY dataset. Our experiments reveal that adding a third modality significantly enhances the performances of CoMM and InfMasking. CoMM as a strong baseline shows performance gains of 7.1% and 1.5% on Vision&Touch and UR-FUNNY, respectively. Although InfMasking's performance gain from adding the third modality is relatively modest compared to CoMM, it still matches CoMM's performance on the Vision&Touch dataset. On the UR-FUNNY dataset, InfMasking achieves the best result (+0.8%).

4.2.3 Experiments with 2 Modalities on Multimodal IMDb

Multimodal IMDb(MM-IMDb) [2] is a real-world multimodal, multi-label dataset designed for movie genre classification. It poses two major challenges: significant class imbalance with genres such as comedy and drama dominating the label distribution, and substantial semantic discrepancy between visual (poster) and textual (plot's description) modalities. Since genre prediction based on a single modality is often unreliable and the combination of both modalities can perform better [2], this underscores the need for effective modeling of multimodal interactions. We select both single-modal and multi-modal as baselines. For unimodal, we choose SimCLR (image-only) [9] and CLIP (pretrained on image-text pairs) [38]. For multimodal, we include CLIP, SLIP [35], and CoMM.

Table 4: Linear probing F1-score (weighted and macro) (in %) for MM-IMDB. △ indicates further training on unlabeded data. ♣ denotes results are from [12].

Model	*Modalities*	*weighted-f1*↑	*macro-f1*↑
SimCLR♣△ [9]	V	$40.35_{\pm 0.23}$	$27.99_{\pm 0.33}$
CLIP♣ [38]	V	51.5	40.8
	L	51.0	43.0
	V+L	58.9	50.9
SLIP♣△ [35]	V+L	$56.54_{\pm 0.19}$	$47.35_{\pm 0.27}$
CLIP♣△ [38]	V+L	$54.49_{\pm 0.19}$	$44.94_{\pm 0.30}$
CoMM(CLIP backbone)	V+L	$61.29_{\pm 0.73}$	$53.79_{\pm 0.22}$
InfMasking(ours, CLIP backbone)	V+L	$\mathbf{62.60}_{\pm 0.26}$	$\mathbf{55.93}_{\pm 0.19}$

Tab. 4 summarizes the experimental results. Models trained on both modalities consistently outperformed their single-modality counterparts, further validating the importance of optimizing multimodal representation learning. InfMasking achieves the best overall performance, improving upon CoMM by 1.31% in weighted F1-score and 2.14% in macro F1-score. It is also worth noting that CLIP with its original public weights achieves 58.9% on weighted F1-score, outperforming CLIP fine-tuned on MM-IMDb (54.59%). This suggests that redundant information learning may not always benefit complex tasks such as genre prediction, which require complementary modality alignment [12]. These results demonstrate the robustness and generalizability of InfMasking in handling imbalanced, semantically heterogeneous, and multi-label multimodal classification tasks.

5 Ablation Studies

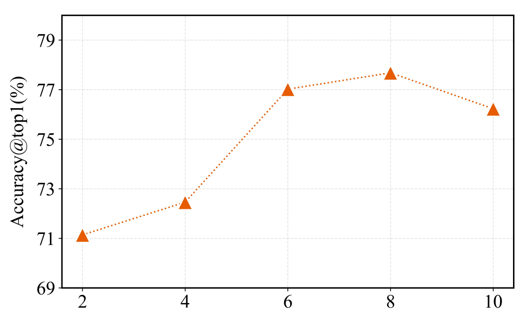
To examine the effectiveness of the design of InfMasking, we conduct comprehensive ablation studies on the bimodal Trifeature dataset focusing on three critical components: the loss function formulation, the optimal number of masked views, and the masking ratio parameter.

Table 5: Linear probing accuracy (in %) of redundancy R, uniqueness U and synergy S on Trifeature Dataset under different combinations of loss functions. λ_1, λ_2, and λ_3 denote the weights for $\mathcal{L}_{mask}$, $\sum_i \mathcal{L}_i$, and $\mathcal{L}$, respectively, where $\mathcal{L}$ and $\sum_i \mathcal{L}_i$ correspond to the first and second terms in Eq. 1.

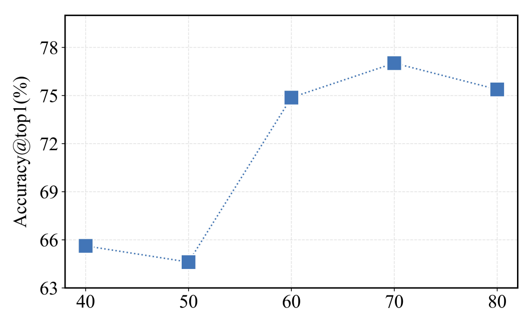
loss weights			R	U_1	U_2	S	*Average*
λ_1	λ_2	λ_3					
0	0	1	$95.8_{\pm 1.91}$	$85.9_{\pm 2.11}$	$83.8_{\pm 2.97}$	$58.7_{\pm 7.11}$	80.1
0	1	1	$99.9_{\pm 0.06}$	$87.1_{\pm 3.31}$	$86.5_{\pm 2.60}$	$71.4_{\pm 3.47}$	86.0
1	1	0	$\mathbf{99.9}_{\pm 0.08}$	$\mathbf{90.7}_{\pm 2.10}$	$\mathbf{91.4}_{\pm 3.03}$	$69.2_{\pm 6.20}$	87.8
1	1	1	$99.9_{\pm 0.09}$	$90.3_{\pm 1.52}$	$90.8_{\pm 2.88}$	$\mathbf{77.0}_{\pm 4.22}$	**89.5**

Loss function. We conducted an ablation study on the Trifeature dataset to evaluate different loss combinations for capturing multimodal interactions. As shown in Tab. 5, the full objective ($\lambda_1 = \lambda_2 = \lambda_3 = 1$, InfMasking) achieves the highest synergy at 77.0% while maintaining balanced performance across other metrics. Using only CoMM loss ($\lambda_1 = 0$, $\lambda_2 = 1$, $\lambda_3 = 1$) yields 71.4% synergy, while using only $\mathcal{L}$ ($\lambda_1 = 0$, $\lambda_2 = 0$, $\lambda_3 = 1$) further decreases to 58.7%, indicating that CoMM loss alone is insufficient without the view diversity from masking. Excluding the $\mathcal{L}$(*i.e.*, $\hat{I}_{\text{NCE}}(Z', Z'')$) drops synergy to 69.2%, despite marginal improvements in redundancy and uniqueness scores. As noted in CoMM [12], minimizing $\mathcal{L}$ enables the model to capture all information terms, albeit at a slower rate. When this loss is removed, the model learns redundancy and uniqueness more efficiently, achieving higher scores within the same epoch, but at the cost of diminished synergy performance.

Number of masked views. According to Section 3.2, increasing the number of masked views yields a closer approximation to $\mathbb{E}_{\text{mask}}\left[\hat{I}_{\text{NCE}}(Z'_{\text{mask}}, Z')\right]$ and $\mathbb{E}_{\text{mask}}\left[\hat{I}_{\text{NCE}}(Z''_{\text{mask}}, Z'')\right]$, albeit at a higher computational cost. As observed in Fig. 2(a), the synergy score improves progressively with an increasing number of views. Notably, performance is sufficiently robust when the number is in the range of $[6, 10]$, demonstrating practical feasibility for GPU implementation.



(a) Number of masked views.



(b) Ratio of masking (%).

Figure 2: Synergy accuracy changes with different masked setting on Trifeature datasets.

Masking ratio. Fig. 2(b) illustrates the impact of varying the masking ratio. At lower ratios ($\leq 50\%$), although the model can capture synergistic information, the overall performance remains unsatisfactory. In contrast, maintaining a higher masking ratio enables the model to generate superior multimodal representations that effectively leverage complementary information across modalities. Furthermore, a higher masking ratio can also provide a greater speedup benefit [19].

6 Related Work

Multimodal learning. Multimodal learning integrates diverse data sources—such as text, image, audio, and tactile inputs—to enhance information understanding across modalities [31, 36, 17, 32, 10]. Traditional approaches rely on simple fusion techniques like feature concatenation [11] or modality-specific prediction averaging [14]. Transformer architectures revolutionized this field through dynamic cross-modal attention mechanisms [42, 50]. Contemporary approaches typically follow a two-stage framework: training specialized encoders for each modality, then projecting these representations into a unified embedding space [4, 12]. This paradigm has been applied across representation learning [5, 38], cross-modal alignment [28, 21], and generative modeling [1, 39].

Self-Supervised multimodal representation learning. Self-supervised learning generates supervision signals from data's inherent structure [13, 24, 48, 47]. In multimodal contexts, these approaches leverage cross-modal correspondences [53, 45]. Prior methods have explored generative approaches, such as reconstructing one modality from another [39], and masked prediction strategies for joint modality modeling [3]. Contrastive learning has emerged as particularly effective for multimodal representation learning [15, 38], constructing positive pairs through data augmentation [9] and introducing both intra-modal and cross-modal objectives [34]. Some approaches incorporate regularization terms to align representations across modalities [44].

Contrastive Multimodal Interactions. Contrastive multimodal approaches [38, 23] optimize cross-modal contrastive loss but emphasize redundant information while neglecting unique and synergistic information requiring joint consideration, with FactorCL [29] addressing this through explicit modeling despite practical implementation challenges. CoMM [12] advances the field using multimodal augmentations and information theory-grounded losses to capture various interaction patterns, though enhancing synergistic interactions remains challenging.

7 Conclusion

This paper introduces InfMasking, a contrastive method that effectively captures synergistic information in multimodal representation learning by stochastically occluding features during fusion and aligning representations through mutual information maximization. We derive a computationally efficient approximation for infinite masking patterns and demonstrate that our approach not only enhances synergistic information extraction in controlled settings but also achieves state-of-the-art performance across seven diverse multimodal benchmarks. InfMasking has some limitations. It lacks a rigorous theoretical framework to systematically analyze the mechanisms of synergistic interactions. Future research will prioritize developing comprehensive theoretical foundations to formally characterize and measure the synergistic information.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U24A20323), the Major Science and Technology Special Project of the Sichuan Provincial Department of Science and Technology (Grant No. 2024ZDZX0002), the Sichuan Provincial Innovation Group Project (Grant No. 2024NSFTD0054), Fundamental Research Funds for the Central Universities (JBK202511081), Blockchain Research Center of China, Natural Science Foundation of China under Grant No. 62502397, National Natural Science Foundation of China (NSFC) (72471197), and the Sichuan Provincial Philosophy and Social Science Fund (Grant No. SCJJ25ND091) . We thank anonymous reviewers for their insightful feedback and Benoit Dufumier, the first author of CoMM [12], for his contributions to open sourcing the code.

References

[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 248–265. Springer, 2020.

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

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 348–367, 2022.

[4] 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.

[5] George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning. *arXiv preprint arXiv:2011.07191*, 2020.

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

[7] Qi Cai, Yu Wang, Yingwei Pan, Ting Yao, and Tao Mei. Joint contrastive learning with infinite possibilities. *Advances in Neural Information Processing Systems*, 33:12638–12648, 2020.

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

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

[10] Yong Dai, Duyu Tang, Liangxin Liu, Minghuan Tan, Cong Zhou, Jingquan Wang, Zhangyin Feng, Fan Zhang, Xueyu Hu, and Shuming Shi. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *arXiv preprint arXiv:2205.06126*, 2022.

[11] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.

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

[13] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022.

[14] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. 2008.

[15] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[16] GroupLens research. MovieLens dataset, 2015.

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

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

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[21] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.

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

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pages 4904–4916, 2021.

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

[25] 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.

[26] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[27] 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.

[28] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16420–16429, 2022.

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

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

[31] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

[32] Che Liu, Yingji Zhang, Dong Zhang, Weijie Zhang, Chenggong Gong, Haohan Li, Yu Lu, Shilin Zhou, Yue Lu, Ziliang Gan, et al. Nexus: An omni-perceptive and-interactive model for language, audio, and vision. *arXiv preprint arXiv:2503.01879*, 2025.

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

[34] Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. ConGeo: Robust cross-view geo-localization across ground view variations. *arXiv preprint arXiv:2403.13965*, 2024.

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

[36] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, pages 689–696, 2011.

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

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

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831, 2021.

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

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[42] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

[43] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations (ICLR)*, 2021.

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

[45] Shuo Wang, Shunyang Huang, Jinghui Yuan, Zhixiang Shen, and Zhao Kang. Cooperation of experts: Fusing heterogeneous information with large margin. *arXiv preprint arXiv:2505.20853*, 2025.

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

[47] Liangjian Wen, Xiasi Wang, Jianzhuang Liu, and Zenglin Xu. MVEB: self-supervised learning with multi-view entropy bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(9):6097–6108, 2024.

[48] Liangjian Wen, Yiji Zhou, Lirong He, Mingyuan Zhou, and Zenglin Xu. Mutual information gradient estimation for representation learning. In *International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30, 2020*.

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

[50] 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.

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

[52] 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.

[53] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2024.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We mainly claim 'synergistic information constitutes the fundamental value proposition of multimodal representation', and we propose a contrastive synergistic information method designed to enhance synergistic information through an Infinite strategy. To verify the effectiveness, we tested on real-world datasets across diverse domains (healthcare, robotics) and data types (image, text, audio). InfMasking achieves state-of-the-art results on seven multimodal tasks involving two or three modalities.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We claim the need for further application to pertaining and post-training, which will benefit more, which is presented in the conclusion section.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

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

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: This study introduces **InfMasking**, a multimodal contrastive interaction method that enhances synergistic Information by utilizing Infinite Masking. InfMasking strategically masks substantial portions of features from each modality during the fusion process, and then maximizes the mutual information between these masked fused features and their unmasked counterparts. Through controlled experiments, we demonstrate that InfMasking effectively enhances synergistic information between modalities. In evaluations on large-scale real-world datasets, InfMasking achieves state-of-the-art performance across seven benchmarks. Detailed analysis and results can be found in Sec. 4. We also present the key hyperparameters in Appendix A.

   Guidelines:

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

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: After the anonymity period, we will open-source our training code and training data.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: As shown in Appendix A, we provide comprehensive details of both training and evaluation procedures.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

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

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix A provides detailed information about the computational resources utilized in our experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This study develops a representation learning model (InfMasking) using publicly available datasets (compliant with their respective licenses) and does not involve human participants or personally identifiable information, thus requiring no IRB approval. We have provided detailed experimental descriptions and plan to open-source the training code and data after the anonymity period to ensure reproducibility. The paper discusses the method's limitations and potential societal impacts, including bias risks and mitigation strategies. However, due to the absence of a formal IRB process in the research location, we conducted an internal peer review to address ethical considerations, as suggested by the NeurIPS guidelines. Additionally, local data privacy regulations prevent immediate data sharing during the anonymity period, but we commit to full transparency post-anonymity. After reviewing the NeurIPS Code of Ethics, we confirm that the study complies with most requirements, with the noted exceptions being addressed through alternative measures.

Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Please refer to Appendix C.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: We utilize the pre-processed datasets provided by MultiBench [30], which have been anonymized to safeguard personal privacy. For the MM-IMDb experiments, we employ publicly available open-source pre-trained models that adhere to established security guidelines.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

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

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper introduces new assets, including the training code and dataset used for the InfMasking representation learning model. These assets are thoroughly documented in the experimental section, which provides detailed descriptions of the model architecture, training procedures, and dataset characteristics, along with citations to all foundational frameworks and packages used. For the open-source code, we include comprehensive comments explaining the functionality of imported packages and the structure of the foundational code. A structured documentation template, detailing training configurations, dataset licensing (compliant with applicable open-source licenses), and model limitations, will be provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: In this work, the LLM is utilized solely for writing and editing purposes. They do not influence the core methodology, scientific rigor, or the originality of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Experimental Details

Training protocol. All experiments are conducted using five independent runs with random seeds in the range $[42, 46]$. We report the mean and standard deviation of performance metrics (*i.e.*, accuracy, mean squared error) to account for variability across runs. Early stopping based on validation accuracy is systematically applied to prevent overfitting. The best-performing checkpoint on the validation set is selected for final evaluation on the test set.

For dataset-specific encoder architectures, modality-specific data augmentation and latent converters, we follow the same configurations as CoMM [12].

Training details. We use AdamW [33] as the optimizer in all experiments. Detailed hyperparameters are listed in Tab. 6. Following [12] on MM-IMDb, we also use a cosine scheduler with final value 10^{-6} and a warmup over 10 epochs. And all models are trained for 100 epochs except for MM-IMDb which is trained for 70 epochs. All experiments are conducted on a single NVIDIA 4090 GPU with 24GB memory.

Table 6: Hyperparameters for InfMasking. *Masking ratio* is the ratio of masking for each masked view. The V&T CP and V&T EE are the contact prediction and end-effector position prediction tasks on Vision&Touch dataset respectively.

dataset	*learning rate*	*masking ratio*	*number of masked views*
Trifeature	3×10^{-4}	0.7	6
MIMIC	3×10^{-4}	0.8	6
MOSI	1×10^{-3}	0.8	5
UR-FUNNY(2 modalities)	1×10^{-3}	0.5	4
MUSTARD	1×10^{-3}	0.5	5
V&T CP(2 modalities)	1×10^{-4}	0.7	6
V&T EE	1×10^{-4}	0.5	4
MM-IMDb	1×10^{-3}	0.8	4
UR-FUNNY(3 modalities)	1×10^{-3}	0.5	4
V&T CP(3 modalities)	1×10^{-4}	0.8	5

Fusion module configuration. For all experiments involving InfMasking, we employ the fusion module similar to that used in CoMM [12], which operates on a sequence of modality-specific embeddings and is implemented as a Transformer-based encoder layer. Specifically, the architecture consists of multi-head self-attention followed by a feed-forward network, with residual connections and layer normalization. In the bimodal setting, we use a 1-layer Transformer with 8 attention heads, while in the trimodal setting, a 2-layer Transformer with the same number of heads is adopted. In addition, a learnable `[CLS]` token is appended to the input sequence, which serves as a global representation aggregating information across modalities.

B Dataset Details

B.1 Trifeature

The Trifeature dataset [22] is designed to investigate the properties of visual neural networks and comprises three distinct features: shape, color, and texture. Each feature consists of 10 categories, resulting in 1,000 unique combinations. Of these, 800 are used for training and 200 for testing. Each training combination is instantiated three times with random rotations applied to both shape and texture components. Shapes are rendered within a 128×128 bounding box, with rotation angles uniformly sampled from $[-45°, 45°]$, and then randomly placed within a 224×224 image canvas while ensuring full visibility. Texture and color are independently applied in the same manner. Image pairs are constructed from these instances, resulting in 10,000 training pairs and 4,096 test pairs, both sampled from the same underlying distribution.

B.2 Multibench

According to [30], all datasets below have been pre-processed to ensure the removal of any personally identifiable information and to safeguard user privacy (some datasets don't include any personal information, *e.g.* Vision&Touch and MM-IMDb).

- **MIMIC** [25] comprises 53,423 hospital admissions from 38,597 distinct patients, spanning the years 2001 to 2012. It includes two modalities: hourly clinical measurements over a 24-hour period (represented as 12-dimensional vectors, times series modality) and static patient information such as age and gender (represented as 5-dimensional vectors, tabular modalities). The task is a binary classification problem aiming to predict whether a patient belongs to ICD-9(*International Statistical Classification of Diseases and Related Health Problems*) code group 7 (460–519), which is commonly used in studies on disease classification [29].
- **MOSI** [52] consists of 2,199 video clips collected from YouTube, designed for sentiment analysis tasks. Each sample includes video, audio signals, and corresponding text transcriptions. The original annotations range continuously from -3 to 3; following the approach in [29], these labels are binarized into positive and negative classes. The model is trained based on textual and visual modalities.
- **UR-FUNNY** [18] is constructed from 1,866 TED talk videos and comprises 16,514 samples for the task of humor detection. Each sample contains video, audio, and corresponding text transcriptions. The objective is to determine whether a given sequence is humorous, formulated as a binary classification problem. For the bi-modal setting, we use the textual and visual modalities.
- **MUSTARD** [8] is designed for sarcasm detection and is sourced from television shows such as Friends. It contains 690 balanced utterances, each comprising video, audio, and text transcriptions, annotated as either sarcastic or non-sarcastic. In our experiments, we utilize the textual and visual modalities.
- **Vision&Touch** [27] comprises data from robotic manipulation tasks, consisting of 150 trajectories, each with 1,000 time steps. The dataset includes RGB images, depth maps, force measurements, and end-effector positions and velocities. The benchmark tasks are (1) binary classification to predict whether contact will occur in the next step and (2) regression to predict the end-effector position, evaluated using mean squared error (MSE). For the bi-modal setting, we use the visual and proprioceptive modalities.

B.3 MM-IMDb

Multimodal IMDb (MM-IMDb) [2] is designed for movie genre prediction and comprises 25,959 films, each annotated with posters, plot summaries, genre labels, and metadata. Derived from the Movielens 20M dataset [16], this benchmark focuses on 23-way multi-label classification. In our experiments, we utilize the image modality (movie posters) and the text modality (plot summaries). While MM-IMDb is part of the Multibench benchmark [30], we present it separately in our experiments, as our model is trained directly on the raw data instead of relying on the pre-processed features offered by Multibench.

C Broader Impact

This study aims to enhance the modelling of cross-modal synergy to generate more informative multimodal representations. To ensure that InfMasking can be deployed responsibly in real-world scenarios, we highlight several key considerations.

Computational complexity. InfMasking introduces an infinite masking strategy that maximizes the mutual information between masked fused views and their unmasked counterparts, strengthening the complementarity of different modalities. However, this method inevitably increases GPU memory usage, as each additional masked view amplifies the memory footprint. We encourage future work to explore lightweight variants that can alleviate the associated computational demands.

Privacy and security. As discussed in Section 4.2.1, the datasets used in this study span multiple domains, including healthcare, sentiment analysis and multimedia. According to [30], all instances containing personal information have been rigorously anonymized and de-identified. And the Vision&Touch and MM-IMDb datasets do not contain any personally identifiable information. All experiments are conducted using irreversible, pre-extracted features, except for MM-IMDb, which is processed directly from raw data; no raw or reconstructable user data is accessed, thereby minimizing privacy risks.

Future work. Future research will focus on establishing rigorous theoretical frameworks to quantify and formally characterize the synergistic information extracted by InfMasking. Such frameworks would provide mathematical guarantees on information preservation while elucidating the fundamental limits of multimodal representation learning. Additionally, we aim to develop adaptive masking strategies that dynamically optimize masking patterns based on task requirements and modality-specific characteristics, potentially employing reinforcement learning to fine-tune these configurations. These advancements would significantly enhance our capacity to model complex synergistic relationships in multimodal data, advancing the field toward more generalizable multimodal intelligence.

D Additional Experiments

D.1 Difference with MAE

Masked Autoencoders(MAE) [19] achieve self-supervised learning through reconstruction of masked image patches, which consist of two parts. For the encoder, it encodes randomly masked image patches into latent features. The decoder is trained to predict the masked patches using reconstruction loss, thereby enhancing semantic relationships between them in a single-modal (vision) setting. InfMasking adapts and extends this masking paradigm to a contrastive multimodal context. It focuses on aligning and extracting synergistic information from multimodal tokens (*e.g.*, features from text, images, audio, or tabular data) through infinite masking, emphasizing cross-modal interactions like redundancy, uniqueness, and synergy.

Table 7: Linear probing accuracy (%) on three datasets from MultiBench [30] for MAE, CoMM, and InfMasking models.

Dataset	MIMIC	UR-FUNNY	MOSI	average
CoMM	$66.4_{\pm 0.41}$	$63.3_{\pm 0.51}$	$63.7_{\pm 2.5}$	64.47
MAE	$67.4_{\pm 0.3}$	$62.5_{\pm 1.43}$	$65.4_{\pm 1.6}$	65.1
InfMasking	$68.1_{\pm 0.42}$	$64.3_{\pm 0.9}$	$69.0_{\pm 1.2}$	67.12

Unlike MAE, our masking approach does not mask the raw input of each modality but rather masks the features of each modality before fusion. Furthermore, we aim to maximize mutual information between masked and unmasked multimodal representations without reconstruction. It derives a lower bound approximation for the InfMasking loss assuming Gaussian distributions for masked features, making it computationally feasible for infinite views. This makes InfMasking a natural evolution for handling diverse modalities, addressing limitations in traditional contrastive learning (*e.g.*, over-reliance on multiview redundancy) while preserving MAE's core idea of using masking to create challenging, informative views.

We further compare our InfMasking loss with MAE reconstruction loss on multiple datasets from MultiBench [30]. The results are illustrated in Tab. 7. Under the same experimental conditions, we randomly mask tokens across modalities, encoder forward pass with masked iuput, and decoder-based reconstruction focused solely on masked tokens (using MSE loss averaged over masked positions). This creates a generative baseline analogous to MAE but extended to multimodal tokens. The MAE variant replaces the InfMasking loss component terms in Eq. (2) with reconstruction loss.

D.2 Ablation Studies of Data Augmentation on Trifeature Datasets

The InfoMin Principle [41] plays a pivotal role in self-supervised learning. It demonstrates that data augmentation is an effective strategy for adhering to this principle, as stronger data augmentations reduce mutual information to an optimal level. In our work, we adopt the same settings for modality-specific data augmentation as outlined in CoMM [12]. To extend this investigation, we further explore the influence of data augmentation strategies on the bimodal Trifeature dataset [22].

Table 8: Impact of data augmentation on linear probing accuracy (%) for multimodal interactions. The term "All" refers to SimCLR [9] augmentations. InfMasking applies "All" augmentations to both modalities, consistent with CoMM.

Augmentations		R	U_1	U_2	S	Average
Modality 1	Modality 2					
{All}	∅	$99.78_{\pm 0.08}$	$85.28_{\pm 2.88}$	$49.89_{\pm 8.73}$	$50.0_{\pm 0.0}$	71.24
∅	{All}	$99.85_{\pm 0.06}$	$49.08_{\pm 2.65}$	$87.44_{\pm 3.59}$	$50.0_{\pm 0.0}$	71.59
{All} \{crop}	{All}	$97.70_{\pm 0.84}$	$58.07_{\pm 2.68}$	$87.15_{\pm 3.80}$	$50.0_{\pm 0.0}$	73.23
{All}	{All} \{crop}	$96.91_{\pm 1.99}$	$85.42_{\pm 4.01}$	$57.85_{\pm 6.63}$	$50.0_{\pm 0.0}$	72.54
InfMasking		$99.86_{\pm 0.10}$	$90.30_{\pm 1.52}$	$90.80_{\pm 2.88}$	$77.02_{\pm 4.22}$	89.5

As shown in Tab. 8, omitting data augmentation leads to a significant degradation in model performance, particularly in uniqueness. Notably, cropping as a critical transformation in self-supervised learning for vision tasks [9, 20], is vital for learning synergistic representations in the Trifeature dataset. When modality-specific cropping augmentation is omitted, the model struggles to capture the uniqueness of the corresponding modality, resulting in an inability to effectively learn synergy.

E Analysis of Gaussian Approximation Assumption via Visualization

Based on the theoretical framework of InfMasking discussed in Sec. 3.2, we further analyze the robustness of Gaussian approximation assumption through visualization.

We employ dimensionality reduction to project the high-dimensional embeddings of multimodal features and their masked view features from the Trifeature dataset into a two-dimensional space, which are visualized using t-SNE. As shown in Fig. 3, the masked embeddings cluster around a central point in the projected space. Notably, this central point aligns closely with the multimodal features, indicating that despite the perturbations introduced by masking, the masked embeddings inherently preserve core aspects of the synergistic semantic nature. It uncovers the stability of synergistic integration: even when parts of modalities are obscured, the fused representations converge toward a shared semantic manifold, reflecting the emergent properties that arise from modal complementarity rather than redundancy alone. Furthermore, the dispersion of masked embeddings similarly indicates semantic differentiation within the ambient embedding space. The observed variance highlights subtle nuances in how masking affects the differences of representations combined from masked view features, potentially corresponding to variations in semantic granularity—such as implicit biases or contextual implications that become apparent only through joint modal analysis.

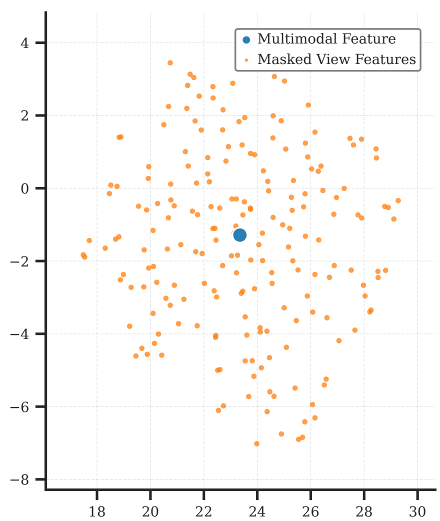


Figure 3: Visualization of the distribution of multimodal fusion embeddings and its masked counterpars.

F Pseudo-Code

Algorithm 1 outlines the training procedure of InfMasking, formulated in the general case with n modalities (*e.g.*, image, text, audio, *etc*).

The key input components are as follows: $\mathcal{T}^\star$ denotes a set of label-preserving transformations used for data augmentation. The fusion transformer g integrates latent features from diverse modalities. The masked view number M' indicates how many masked instances are generated per modality. The random masking operator $\mathcal{M}$ stochastically obscures portions of the embedding features. And the temperature parameter τ controls the sharpness of the total loss.

Algorithm 1 Multimodal contrastive interaction learning with InfMasking

Input: Multi-modal dataset $\{X_1, X_2, ..., X_n\}$, label-preserving transformations $\mathcal{T}^{\star}$, set of projection transformations $\mathcal{P} = \{p_1, \ldots, p_n\}$, batch size N, masked view number M', uni-modal encoders $(f_i)_{i\in[1..n]}$, fusion transformer g, random mask operator $\mathcal{M}$, temperature parameter τ.

for sampled mini-batch $\{\mathbf{x}_k\}_{k\in[1..N]} = (\mathbf{x}_k^1, ..., \mathbf{x}_k^n)_{k\in[1..N]}$ **do**

  for $k \in [1..N]$ **do**

    draw $t', t'' \sim \mathcal{T}^{\star}$

    $\mathbf{x}'_i, \mathbf{x}''_i \leftarrow t'(\mathbf{x}_i), t''(\mathbf{x}_i)$

    for $j \in [1..M']$ **do**

      $\mathbf{z}'^{j}_{mask} \leftarrow g\left(\mathcal{M}(f_1(\mathbf{x}'^{1}_{k})), ..., \mathcal{M}(f_n(\mathbf{x}'^{n}_{k}))\right)$

      $\mathbf{z}''^{j}_{mask} \leftarrow g\left(\mathcal{M}(f_1(\mathbf{x}''^{1}_{k})), ..., \mathcal{M}(f_n(\mathbf{x}''^{n}_{k}))\right)$

    end for

    $\mathbf{z}'_k \leftarrow g(f_1(\mathbf{x}'^{1}_{k}), ..., f_n(\mathbf{x}'^{n}_{k}))$

    $\mathbf{z}''_k \leftarrow g(f_1(\mathbf{x}''^{1}_{k}), ..., f_n(\mathbf{x}''^{n}_{k}))$

    for $i \in [1..n]$ **do**

      $\mathbf{x}_k^i \leftarrow p_i(\mathbf{x}_k)$

      $\mathbf{z}_k^i \leftarrow g(f_i(\mathbf{x}_k^i))$

    end for

  end for

  $$\mathcal{L}_{\text{InfMasking}} \leftarrow -\frac{1}{M'}\sum_{k=1}^{M'}\Bigg[\mathbb{E}_{z'^{k}_{mask}, z'_{\text{pos}} \sim p(Z'^{k}_{mask}, Z')}\left[\log \frac{\exp(z'^{k\,T}_{mask} z'_{\text{pos}}/\tau)}{\exp(z'^{k\,T}_{mask} z'_{\text{pos}}/\tau) + \sum_{z'_{\text{neg}}} \exp(z'^{T} z'_{neg}/\tau)}\right]$$

  $$+\mathbb{E}_{z''^{k}_{mask}, z''_{\text{pos}} \sim p(Z''^{k}_{mask}, Z'')}\left[\log \frac{\exp(z''^{k\,T}_{mask} z''_{\text{pos}}/\tau)}{\exp(z''^{k\,T}_{mask} z''_{\text{pos}}/\tau) + \sum_{z''_{\text{neg}}} \exp(z''^{T} z''_{neg}/\tau)}\right]\Bigg]$$

  for $i \in [1..n]$ **do**

    $$\mathcal{L}_i \leftarrow -\Bigg[\mathbb{E}_{z_i, z'_{\text{pos}} \sim p(Z_i, Z')}\left[\log \frac{\exp(z_i^T z'_{\text{pos}}/\tau)}{\exp(z_i^T z'_{\text{pos}}/\tau) + \sum_{z'_{\text{neg}}} \exp(z_i^T z'_{\text{neg}}/\tau)}\right]$$

    $$+\mathbb{E}_{z_i, z''_{\text{pos}} \sim p(Z_i, Z'')}\left[\log \frac{\exp(z_i^T z''_{\text{pos}}/\tau)}{\exp(z_i^T z''_{\text{pos}}/\tau) + \sum_{z''_{\text{neg}}} \exp(z_i^T z''_{\text{neg}}/\tau)}\right]\Bigg]$$

  end for

  $$\mathcal{L} \leftarrow -\mathbb{E}_{z', z''_{\text{pos}} \sim p(Z', Z'')}\left[\log \frac{\exp(z'^{T} z''_{\text{pos}}/\tau)}{\exp(z'^{T} z''_{\text{pos}}/\tau) + \sum_{z''_{\text{neg}}} \exp(z'^{T}, z''_{\text{neg}}/\tau)}\right]$$

  $\mathcal{L}_{\text{Total loss}} \leftarrow \mathcal{L} + \sum_{i=1}^{n} \mathcal{L}_i + \mathcal{L}_{\text{InfMasking}}$

  update $(f_i)_{i\in[1..n]}, \mathcal{M}, g$ to minimize $\mathcal{L}_{\text{Total loss}}$

end for

return $(f_i)_{i\in[1..n]}, g$

G Proof

Proof 1 (lemma 3)

$$\begin{aligned}
&\mathbb{E}_{mask}[\hat{I}_{NCE}(Z'_{mask}, Z')] && (8)\\
&= \mathbb{E}_{mask}[\mathbb{E}_{z'\sim p(Z')}\left[\log \frac{\exp(z'^T z'_{mask}/\tau)}{\exp(z'^T z'_{mask}/\tau) + \sum_{z'_{neg}} \exp(z'^T z'_{neg}/\tau)}\right]] && (9)\\
&= \mathbb{E}_{z'\sim p(Z')}[\mathbb{E}_{mask}\left[\log \frac{\exp(z'^T z'_{mask}/\tau)}{\exp(z'^T z'_{mask}/\tau) + \sum_{z'_{neg}} \exp(z'^T z'_{neg}/\tau)}\right]] && (10)\\
&= \mathbb{E}_{z'\sim p(Z')}[\mathbb{E}_{mask}\left[(z'^T z'_{mask}/\tau) - \log[\exp(z'^T z'_{mask}/\tau) + \sum_{z'_{neg}} \exp(z'^T z'_{neg}/\tau)]\right]] && (11)\\
&\geq \mathbb{E}_{z'\sim p(Z')}\left[z'^T \mathbb{E}_{mask}[z'_{mask}]/\tau - \log \mathbb{E}_{mask}[\exp(z'^T z'_{mask}/\tau) + \sum_{z'_{neg}} \exp(z'^T z'_{neg}/\tau)]\right] && (12)
\end{aligned}$$

The inequality Eq.(12) merges from the application of Jensen inequality on concave functions i.e., $\mathbb{E}_x \log(X) \leq \log \mathbb{E}_x[X]$. z'_{mask} denotes the integrated representation derived from the fusion of all masked features across diverse modalities via the Transformer architecture.

Lemma 4 *Consider a random variable $\boldsymbol{x}$ that follows a multivariate Gaussian distribution, denoted as $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^n$ represents the mean vector and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ is the covariance matrix The moment generating function (MGF) of this random variable is given by the following expression:*

$$\mathbb{E}_{\boldsymbol{x}}\left[e^{\boldsymbol{a}^T \boldsymbol{x}}\right] = e^{\boldsymbol{a}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a}}, \tag{13}$$

where $\boldsymbol{a} \in \mathbb{R}^n$ is an arbitrary constant vector.

$z'^T z'_{mask}/\tau$ According to Lemma 4, we can derive the MGF of the inequality Eq.(7) as follows:

$$\begin{aligned}
&\mathbb{E}_{mask}[\hat{I}_{NCE}(Z'_{mask}, Z')] && (14)\\
&\geq \mathbb{E}_{z'\sim p(Z')}\left[z'^T \boldsymbol{\mu}_{z'_{mask}}/\tau - \log[\exp(z'^T \boldsymbol{\mu}_{z'_{mask}}/\tau + \frac{1}{2\tau^2} z'^T \boldsymbol{\Sigma}_{z'_{mask}} z) + \sum_{z'_{neg}} \exp(z'^T z'_{neg}/\tau)]\right] && (15)
\end{aligned}$$