Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Robult: Leveraging Redundancy and Modality-Specific Features for Robust Multimodal Learning

Duy A. Nguyen[1,3], **Abhi Kamboj**[2], **Minh N. Do**[2,3]
[1]Siebel School of Computing and Data Science, UIUC, US
[2]Department of Electrical and Computer Engineering, UIUC, US
[3]VinUni-Illinois Smart Health Center, VinUniversity, Vietnam
{duyan2, akamboj2, minhdo}@illinois.edu

Abstract

Addressing missing modalities and limited labeled data is crucial for advancing robust multimodal learning. We propose **Robult**, a scalable framework designed to mitigate these challenges by preserving modality-specific information and leveraging redundancy through a novel information-theoretic approach. Robult optimizes two core objectives: (1) a soft Positive-Unlabeled (PU) contrastive loss that maximizes task-relevant feature alignment while effectively utilizing limited labeled data in semi-supervised settings, and (2) a latent reconstruction loss that ensures unique modality-specific information is retained. These strategies, embedded within a modular design, enhance performance across various downstream tasks and ensure resilience to incomplete modalities during inference. Experimental results across diverse datasets validate that Robult achieves superior performance over existing approaches in both semi-supervised learning and missing modality contexts. Furthermore, its lightweight design promotes scalability and seamless integration with existing architectures, making it suitable for real-world multimodal applications.

1 Introduction

Motivation: In the Big Data era, multimodal learning significantly improves data exploitation, outperforming single-modality approaches [Huang *et al.*, 2021]. However, most existing methods [Daunhawer *et al.*, 2023; Peng *et al.*, 2022] operate under idealized assumptions: fully labeled training datasets and consistently available modalities during evaluation. In practice, the challenges of missing modalities and semi-supervised learning often coexist, yet current research typically addresses these problems in isolation. For example, missing modalities may arise when autonomous vehicles lose sensor inputs due to environmental obstructions or when medical diagnostics in resource-limited settings lack access to all imaging modalities. Simultaneously, the scarcity of labeled data across domains remains a critical bottleneck, particularly in multimodal contexts where individual modalities often require specialized labeling. Addressing these challenges independently limits the adaptability of multimodal systems and fails to capture the complexities of real-world deployments. This work uniquely addresses these dual challenges by integrating strategies to handle incomplete modalities and leverage unlabeled data simultaneously. By doing so, our approach enhances both flexibility and applicability, enabling robust performance in diverse and imperfect scenarios. Unlike current literature, which rarely addresses both issues concurrently, this work bridges a critical gap with an innovative and unified solution.

Existing literature: One of the primary challenges in multimodal learning is handling corrupted or missing modalities. Existing methods to address missing modalities typically fall into two categories:

1. **Generative approaches**, such as VAE-based models [Wu and Goodman, 2018], which reconstruct missing modalities. While recent advancements [Feichtenhofer *et al.*, 2022; Woo *et al.*, 2023] demonstrate promising performance, these methods often depend on specific architectures, limiting their flexibility.

2. **Transfer learning methods**, which align latent spaces for cross-modal knowledge transfer [Ma *et al.*, 2022; Lee and Van der Schaar, 2021; Wang *et al.*, 2020]. These methods focus on adaptable training strategies [Chen *et al.*, 2023] but often lack a strong theoretical foundation and are primarily guided by empirical intuition.

Simultaneously, the need for semi-supervised learning arises from practical challenges in labeling raw data, especially in domains where annotations are scarce or labor-intensive to obtain. This challenge is amplified in multimodal learning, as each modality may require distinct expertise for labeling. For example, tasks such as object segmentation across video and lidar data in autonomous driving [Zhang *et al.*, 2022b] or medical segmentation across imaging modalities [Acosta *et al.*, 2022] demand diverse and often non-standardized labeling procedures. Recent advancements in semi-supervised learning, such as knowledge distillation [Su *et al.*, 2021] and pseudo-labeling [Aberdam *et al.*, 2022], have shown promise, but these techniques are often designed for specific applications and struggle to generalize across varied multimodal settings.

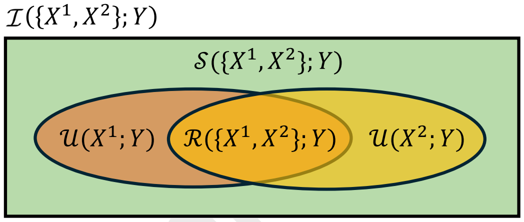


Figure 1. Partial Information Decomposition for 2 input modalities with target variable.

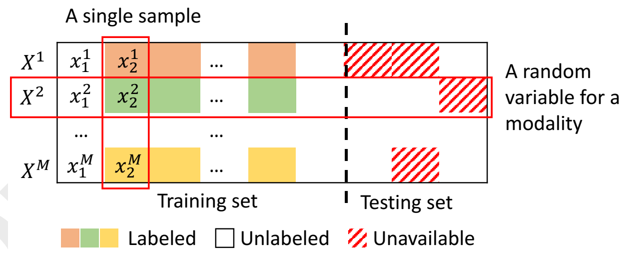


Figure 2. Training/evaluation datasets under investigation.

A detailed discussion of existing literature in both areas is provided in Appendix A.

Proposed approach: The dual challenges of missing modalities and semi-supervised learning remain critical open problems in multimodal research. This work advances transfer learning methods for missing modalities (category 2) that pertain architectural flexibility and dataset compatibility, while introducing a novel design to reduce dependence on labeled data, ensuring robustness and adaptability in real-world scenarios.

Using Partial Information Decomposition [Williams and Beer, 2010], the mutual information provided by an input X with M modalities $(X^1, \ldots, X^M)$ for a given task Y can be decomposed into:

$$\mathcal{I}(\{X^1, \ldots, X^M\}; Y) = \mathcal{R}(\{X^1, \ldots, X^M\}; Y) + \sum_{i=1}^{M} \mathcal{U}(X^i; Y) + \mathcal{S}(\{X^1, \ldots, X^M\}; Y), \tag{1}$$

where $\mathcal{R}$ represents redundancy (shared task-relevant information among M modalities), $\mathcal{U}$ denotes unique information specific to the i^{th} modality, and $\mathcal{S}$ quantifies synergy, the additional knowledge generated through interactions among modalities (Figure 1 illustrates PID with 2 modalities).

In an ideal setting with access to all modalities, a fusion technique would efficiently capture $\mathcal{R}$ and $\mathcal{S}$ to optimize predictions for Y. However, in real-world scenarios where some modalities are unavailable, replicating $\mathcal{S}$ becomes challenging. For example, with two modalities, $Y = \texttt{linear}(X^1, X^2) = \texttt{non\_linear}(X^1)$ demonstrates that synergy ($\mathcal{S}$) improves predictions when both X^1 and X^2 are available. Access to X^1 alone makes the relationship more complex and harder to model with deep networks.

Approaches like knowledge distillation [Chen *et al.*, 2023] and contrastive learning [Radford *et al.*, 2021] aim to address missing modalities by mimicking the representations produced by fused modalities in their absence. From an information-theoretic perspective, these methods focus on replicating redundant information ($\mathcal{R}$) through latent-space alignment. Building on this foundation, we explicitly introduce a mutual information maximization objective (Objective 2.1) to align unimodal and fused representations. This alignment ensures efficient knowledge transfer while minimizing reliance on labeled data using a novel soft-positive pseudo-labeling mechanism that accounts for pseudo-label uncertainty.

While alignment effectively captures $\mathcal{R}$, it can unintentionally diminish unique information ($\mathcal{U}$) from individual modalities. This loss is particularly detrimental in semi-supervised settings or when other modalities are unavailable. To address this, we propose preserving modality-specific information ($\mathcal{U}^i$) via Objective 2.2. This is achieved using a simple yet effective reconstruction procedure in the latent space, universally applicable across modalities. Our results and ablations (Tables 1 and 3) demonstrate how retaining $\mathcal{U}^i$ improves performance.

Together, Objectives 2.1 and 2.2 form the foundation of our semi-supervised multimodal learning method, **Rob**ust **Mult**imodal Pipeline (**Robult**). Robult effectively balances redundancy alignment and unique information preservation, ensuring accuracy and robustness in scenarios with missing modalities. Empirical results (Section 3) and theoretical underpinnings highlight Robult's superior performance compared to existing methods, demonstrating its adaptability to diverse real-world settings.

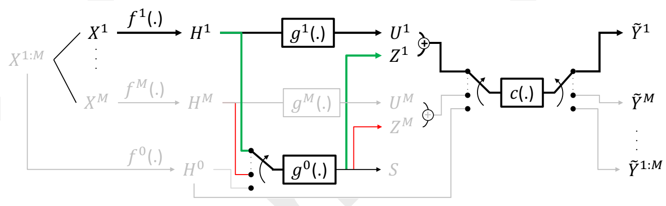
Contributions. Our primary contributions are:

- Jointly addressing the dual challenges of missing modalities and semi-supervised learning.
- Framing two objectives under an information-theoretic perspective and deriving novel loss functions to achieve these goals.
- Introducing a soft Positive-Unlabeled contrastive loss that efficiently utilizes limited labeled data through selective weighting of potential positives.

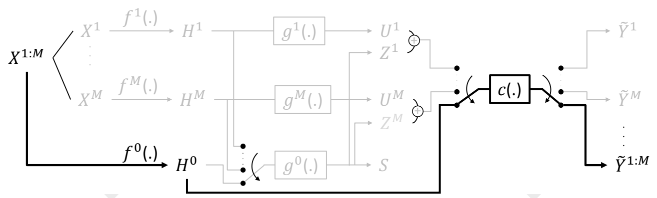
2 Methodology

Training setting: We consider a training scenario where each sample includes all modalities, but only some samples have corresponding labels (Figure 2). Formally, let the training dataset be $\mathcal{D}_{train} = \{(x_1, y_1), \ldots, (x_k, y_k), x_{k+1}, \ldots x_n\}$, where each data point $x_j = (x_j^1, \ldots, x_j^M)$ contains M modalities. The dataset consists of n samples, of which k are labeled ($0 \leq k < n$).

Evaluation setting: To mimic real-world deployment, the evaluation dataset $\mathcal{X}_{test} = \{x_1, \ldots, x_m\}$ consists of samples with potential missing modalities - e.g. $x_j = (x_j^a\text{'s}); \forall a \in$



(a) Inference when some modalities are missing



(b) Inference when all modalities are available

Figure 3. (3a) For each available modality $i \in \{1, \ldots, M\}$, $f^i(.)$ extracts the latent representation h^i_j. The modules $g^i(.)$ and $g^0(.)$ then extract the unique information u^i_j and redundant representation z^i_j, respectively. These are combined and processed by the shared module $c(.)$ to predict $\tilde{y}^i_j$. Late fusion strategies aggregate these outputs into the final prediction $\tilde{y}_j$. (3b) When all modalities $x^{1:M}_j$ are available, $f^0(.)$ extract a fused latent vector h^0_j. This vector is directly passed through $c(.)$ to generate the final output $\tilde{y}^{1:M}_j$.

$\mathcal{A}^M_j \subseteq \{1, \ldots, M\}$. Here, $\mathcal{A}^M_j$ denote the indices of all available modalities for sample x_j (Figure 2).

Notation: For clarity, in the theoretical framework, we denote the input random variables corresponding to the i^{th} modality as X^i (Figure 2). We would refer back to the observation-level notation x^i_j whenever needed.

Figure 3 illustrates our versatile multimodal pipeline named Robult. Robult consists of M modality-specific branches and a fusion branch indexed with zero. For each input variable X^i, it is first projected into a shared latent space:

$$H^i = f^i(X^i), \quad i \in \{1, \ldots, M\}\,.$$

In parallel, a fusion network $f^0(.)$ processes all input modalities jointly to produce a fused latent variable (also reside on the shared latent space):

$$H^0 = f^0(X^{1:M}),$$

where index 0 denotes the joint latent variable using all modalities. This latent-space projection provides two key benefits:

- **Generalization**: Robult supports different combination of modalities and their preferred projection methods/architectures.
- **Efficient Fusion**: Various strategies can be adopted before Robult's main logic to efficiently attain fused representations.

Each modality-specific module $g^i(.)$ extracts unique information of corresponding modality:

$$U^i = g^i(H^i), \quad i \in \{1, \ldots, M\}\,,$$

while the shared module $g^0(.)$ effectively captures the redundant information by processing individual or joint latent variables:

$$Z^i = g^0(H^i), S = g^0(H^0).$$

Objectives: As discussed in Section 1, we draw inspiration from the second category of literature on missing modalities, which primarily leverages knowledge distillation and contrastive loss techniques [Poklukar *et al.*, 2022b]. These methods aim to align unimodal (student) representations, such as Z^i, with the fused (teacher) representation S, enabling unimodal representations to efficiently replicate the redundant information encapsulated in the fused representation. Although this redundant information exists within each modality, extracting it directly from a single modality can be significantly more complex without access to all modalities. For example, encoding object shape from an image may be challenging, whereas a textual description can explicitly provide the same information. Mimicking the fused representation offers a shortcut, simplifying the extraction of redundant information within unimodal representations.

However, these approaches fail in the semi-supervised training setting, owing to 2 emerging challenges:

- (1) The *over reliance on label signals* (e.g. classification clusters) in the alignment process.
- (2) the *diminishing of unique information* contained in different modality after alignment process.

Regarding the former challenge, we directly address the label need of Contrastive Learning with Label-level sampling [Zhang *et al.*, 2022a] by a novel soft Positive-Unlabelled (PU) contrastive loss, together with an adaptive weighting strategy, detailed about which is covered in Section 2.1. Theoretically, this PU contrastive loss corresponds to a mutual information maximization problem between the desired fused representation S and the learned unimodal representation projected into that same latent space Z^i's. Objective 2.1 of our method can be expressed as follow:

Objective 2.1. Aligning S and Z^i by maximizing the mutual information $\mathcal{I}(S, Z^i) \quad (i = \overline{1, M})$.

For the latter challenge, we observe that attempting to only align modalities during training diminishes the unique information provided by each modality, thus the model is losing information that could help inform it when only that specific modality is available. Therefore, a model should ideally benefit by maintaining the redundancy while also explicitly preserving each modality's unique information. This claim is later supported by experimental results and ablations (Section 3 - Table 1). To address the challenge of vanishing modality-specific information from multimodal alignment during training, we emphasize a disentanglement strategy that preserves unique information while still facilitating the redundancy learning process of Objective 2.1. Robult integrates a set of modules $g^i(H^i)$, where $i = 1, \ldots, M$, to produce unique representations U^i for each modality. We aim to preserve the unique information for each modality via the learning of U^i with Objective 2.2 as follows:

Objective 2.2. Learning U^i by minimizing the conditional entropy $\mathcal{H}(H^i|Z^i, U^i) \quad (i = \overline{1, M})$.

During training, all branches are executed and three loss functions update the learned modules. The soft positive unlabeled loss, $\mathcal{L}_{PU}$, maximizes knowledge extraction from the few labeled samples in a batch (Subsection 2.1). The reconstruction loss, $\mathcal{L}_{rec}$ forces each branch to extract unique modality-specific information (Subsection 2.2). The task-specific supervised loss, $\mathcal{L}_{sup}$ is used on the labeled samples across all modules to learn label information (Subsection 2.3).

2.1 Maximizing Mutual Information with Soft Positive-Unlabeled Contrastive Learning

To address the impact of missing modalities, we aim to learn redundant information by aligning the fused latent variable S with unimodal representations Z^i, as formalized in Objective 2.1. This is achieved by maximizing their mutual information $\mathcal{I}(S, Z^i)$. However, direct computation of this quantity is infeasible without access to the joint distribution p_{S,Z^i} or the marginal distributions p_S and p_{Z^i}. Instead, we derive and optimize a lower bound for this mutual information.

Lower Bound Derivation. Let F be a binary random variable indicating whether a pair (s_j, z_k^i) is sampled from the joint distribution p_{S,Z^i} ($F = 1$) or from the product of marginal distributions $p_S \otimes p_{Z^i}$ ($F = 0$). Then, a lower bound for $\mathcal{I}(S, Z^i)$ is expressed as:

$$\begin{aligned}\mathcal{I}(S, Z^i) &\geq -\mathbb{E}_{p_{S,Z^i}} \log v(S, Z^i) \\ &= -\mathbb{E}_{p(S,Z^i|F=1)} \log v(S, Z^i)\end{aligned} \tag{2}$$

where $v(S, Z^i)$ is a non-parametric approximation of $p(F = 1|S, Z^i)$. For a sampled pair (s_j, z_k^i) in a batch of B samples, where s_j is the fused representation for sample j and z_k^i is the i^{th} modality-specific representation for sample k, $v(s_j, z_k^i)$ is defined as:

$$v(s_j, z_k^i) = \frac{\phi(s_j, z_k^i)}{\sum_h^B \phi(s_j, z_h^i)};$$
$$\text{where} \quad \phi(s_j, z_k^i) = exp(\langle s_j; z_k^i \rangle / \tau).$$

The derivation of Result 2 is detailed in Appendix B.2.

Challenges with Sampling. The lower bound in equation 2 relies on expectations under $\mathbb{E}p_{S,Z^i}$, a key source of deviation in existing studies, which presents challenges in practice. Two common sampling strategies include:

1. **Instance-level sampling**, which considers only intra-sample pairs (s_j, z_j^i) within q mini-batch [Radford *et al.*, 2021].
2. **Label-level sampling**, which uses label information to sample inter-sample pairs (s_j, z_k^i) with the same labels, e.g. $y_j = y_k$ [Zhang *et al.*, 2022a].

The first approach risks introducing false negatives, while the second requires fully labeled data, limiting its applicability in semi-supervised settings.

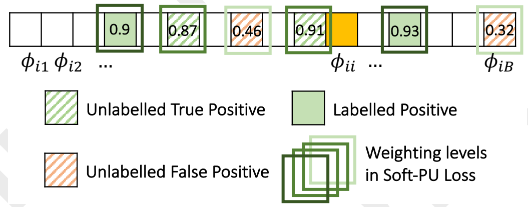


Figure 4. Soft-PU Loss mechanism. Unlabeled positive pairs are identified using soft labels from Robult's classifier. These pairs are re-weighted based on their proximity and the mean proximity of true labeled positive pairs to mitigate false positives.

Soft Positive-Unlabeled (PU) Contrastive Loss. To overcome these limitations, we propose a novel soft Positive-Unlabeled (PU) contrastive loss with adaptive weighting. Let L indicate whether a sampled pair $\langle s_j, z_k^i \rangle$ is labeled ($L = 1$) or not ($L = 0$) (where $L = 1$ if the label information of both samples i and k is known, and $L = 0$ otherwise), The lower bound in equation 2 can then be decomposed as:

$$\begin{aligned}-\mathbb{E}_{p_{S,Z^i}} \log v(S, Z^i) &= -\mathbb{E}_{p(S,Z^i|F=1)} \log v(S, Z^i) \\ &= -p(L{=}1)\mathbb{E}_{p(S,Z^i|F=1,L=1)} \log v(S, Z^i) \\ &\quad - p(L{=}0)\mathbb{E}_{p(S,Z^i|F=1,L=0)} \log v(S, Z^i)\end{aligned} \tag{3}$$

We formulate the two terms in equation 3 as separate loss components: $\mathcal{L}_{lb}$ for labeled data and $\mathcal{L}_{ulb}$ for the unlabeled data.

For labeled samples, let $B_{1,1}^j$ denote the index set of inputs in the batch that share the same class as sample j ($F = 1$) and are labeled ($L = 1$), i.e. $k \in B_{1,1}^j \iff (s_j, z_k^i) \sim p(S, Z^i|F = 1, L = 1)$. The labeled loss is then defined as a NT-Xent-like contrastive loss [Chen *et al.*, 2020]:

$$\mathcal{L}_{lb} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j \in B} -\frac{1}{||B_{1,1}^j||} \sum_{k \in B_{1,1}^j} \log v(s_j, z_k^i). \tag{4}$$

For unlabeled samples, directly sampling from $p(S, Z^i|F = 1, L = 0)$ is challenging due to the absence of label information. To address this, we propose leveraging the output of the Robult classifier to generate soft labels, regularized by adaptive weights calculated dynamically within each mini-batch. This approach effectively balances the influence of soft-labeled pairs and mitigates the limitations of traditional pseudo-labeling methods [Aberdam *et al.*, 2022].

A key challenge during initial training stages is the instability of the Robult classifier, which may produce unreliable outputs and hinder effective filtering of false positives, as illustrated in Figure 4. To counteract this, we adjust the contribution of soft-labeled pairs to the loss function, ensuring robust training. For each anchor sample s_j within a mini-batch, there are labeled positive partners or, in unsupervised scenarios, unimodal representations z_j^i. The average proximity of these labeled partners to s_j serves as a reference for determining "positive" proximity thresholds. Unlabeled positive pairs identified by the Robult classifier are expected to exhibit proximity close to this reference mean. To refine the loss computation, we increase the weight of pairs that align closely with the reference mean and decrease the influence of outliers. This weighting mechanism, implemented using an RBF kernel, allows precise adjustment of each couplet's contribution based on its proximity. By dynamically adapting weights, our method effectively reduces the impact of potential false positives and enhances the overall performance of the final loss function, as demonstrated in Figure 4.

$$\begin{aligned} w_{jk}^i &= RBF(\phi_j^i, \phi(s_j, z_k^i)); \\ \text{where} \quad \phi_j^i &= mean\left\{\phi(s_j, z_{\tilde{k}}^i) | \tilde{k} \in B_{1,1}^j\right\}. \end{aligned} \tag{5}$$

Let $B_{1,0}^j$ denote the index set of inputs in the batch that share the same class as anchor j ($F = 1$) but are not labeled ($L = 0$), i.e. $k \in B_{1,0}^j \iff (s_j, z_k^i) \sim p(S, Z^i | F = 1, L = 0)$. The unlabeled loss is given by:

$$\mathcal{L}_{ulb} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j\in B} -\frac{1}{||B_{1,0}^j||}\sum_{k\in B_{1,0}^j} w_{jk}^i \log v(s_j, z_k^i). \tag{6}$$

The complete soft Positive-Unlabeled (PU) loss is the sum of the labeled and unlabeled components:

$$\mathcal{L}_{PU} = \mathcal{L}_{ulb} + \mathcal{L}_{lb} \tag{7}$$

2.2 Minimizing Conditional Entropy with Latent Reconstruction Error

This section details the procedure for achieving Objective 2.2, which focuses on preserving unique information U^i. Let p_{U^i,Z^i} denote the joint distribution of U^i and Z^i, where $(u_j^i, z_j^i) \sim p_{U^i,Z^i}$ are derived from the corresponding instance h_j^i. Inspired by [Chen *et al.*, 2016], the conditional entropy $\mathcal{H}(H^i|Z^i, U^i)$ is expressed as:

$$\mathcal{H}(H^i|U^i, Z^i) = -\mathbb{E}_{p_{U^i,Z^i}}\left[\mathbb{E}_{p_{H^i|U^i,Z^i}}\left[\log p\left(H^i \mid U^i, Z^i\right)\right]\right] \tag{8}$$

Upper Bound Derivation. Since directly computing $p(H^i \mid U^i, Z^i)$ is challenging, we approximate it using a distribution $q(H^i \mid U^i, Z^i)$. Substituting q into Eq. 8, we derive:

$$\begin{aligned} \mathcal{H}(H^i|Z^i, U^i) &= -\mathbb{E}_{U^i,Z^i}\left[\mathbb{E}_{H^i|U^i,Z^i}\left[\log q(H^i|U^i, Z^i) \cdot \tfrac{p(H^i|U^i,Z^i)}{q(H^i|U^i,Z^i)}\right]\right] \\ &= -\mathbb{E}_{U^i,Z^i}\left[\mathbb{E}_{H^i|U^i,Z^i}[\log q(H^i|U^i, Z^i)] + \mathrm{KL}(p||q)\right] \\ &\leq -\mathbb{E}_{U^i,Z^i}\left[\mathbb{E}_{H^i|U^i,Z^i}[\log q(H^i|U^i, Z^i)]\right] \end{aligned} \tag{9}$$

The last inequality arises because the KL divergence $\mathrm{KL}(p||q)$ is non-negative. Thus, minimizing $\mathcal{H}(H^i|Z^i, U^i)$ reduces to minimizing its Evidence Lower Bound (ELBO)-like [Kingma and Welling, 2014] upper bound using the approximating distribution $q\left(H^i \mid U^i, Z^i\right)$.

Modeling $q\left(H^i \mid U^i, Z^i\right)$ with Latent Reconstruction. We model $q\left(H^i \mid U^i, Z^i\right)$ through a latent reconstruction procedure in the shared latent space. Specifically, we define a reconstruction module $r^i(U^i, Z^i) = \tilde{H}^i$ where $\tilde{H}^i$ approximates H^i. For each pair $(u_j^i, z_j^i) \sim p_{U^i,Z^i}$ generated from h_j^i, the module $r^i(.)$ attempts to reconstruct $\tilde{h}_j^i$ such that it closely resembles h_j^i. The reconstruction loss is formulated as:

$$\mathcal{L}_{rec} = \frac{1}{MB}\sum_{i=1}^{M}\sum_{j=1}^{B} 1 - \langle \tilde{h}_j^i, h_j^i \rangle^2, \tag{10}$$

where $\langle .; . \rangle$ denotes the $L2$-normalized dot product operation, B is the size of mini-batch, and M is the number of modalities.

This reconstruction in latent space is computationally efficient and alleviates the complexity of directly reconstructing raw modality data. Moreover, it generalizes well across various modalities, enhancing the flexibility of Robult. The reconstruction loss $\mathcal{L}_{rec}$ is back-propagated exclusively through the M unimodal branches. This design ensures that the unique information $\mathcal{U}(X^i; Y)$ is preserved through U^i, without interfering with the shared branch's focus on learning redundant information, as discussed in Section 2.1.

2.3 Training strategy

The objectives outlined in Sections 2.1 and 2.2 are optimized through their respective loss functions. Due to the distinct purposes of these objectives, it is advantageous to learn them separately. Specifically, we apply $\mathcal{L}_{rec}$ to guide the learning of $g^i(.)$ ($i = 1, \ldots, M$), while $\mathcal{L}_{PU}$ drives the optimization of $f^i(.), f^0(.)$, and $g^0(.)$. To maximize the utility of labeled data, we incorporate an additional supervised loss $\mathcal{L}_{sup}$. This loss directs the learning process for the entire Robult network and adapts based on the task type: L_1 loss for regression tasks and cross-entropy $\mathcal{L}_{ce}$ for classification tasks. A detailed training procedure, including loss formulations and implementation details, is provided in Appendix B.3.

3 Experimental Results

3.1 Datasets and Metrics

Dataset: We conduct experiments on the following datasets.

- *CMU-MOSI* [Zadeh *et al.*, 2016] & *CMU-MOSEI* [Zadeh *et al.*, 2018b]: Containing three modalities - text, audio, and visual, supporting sentiment analysis and emotion recognition tasks. Each video is labeled on a scale from -3 (negative) to 3 (positive) sentiment.
- *MM-IMDb* [Arevalo *et al.*, 2017]: Containing image and text modalities, serving genre classification task - which involves multi-label classification as a movie has several genres.
- *UPMC Food-101* [Wang *et al.*, 2015]: Containing two modalities, text and images, this dataset is a classification dataset consisting of 101 food categories.
- *Hateful Memes* [Kiela *et al.*, 2020]: Containing text and image modalities, this dataset aims at identifying hate

speech in memes. This dataset includes challenging examples that are similar to hateful ones but are actually harmless.

Metrics: For sentiment analysis related to CMU-MOSI and CMU-MOSEI datasets, we adopt mean absolute error (MAE), correlation (Cor), binary accuracy, and F1 score, following [Poklukar *et al.*, 2022b; Tsai *et al.*, 2018]. Here, binary categories determine positive sentiment scores (> 0) or negative ones (< 0). For the evaluation of the three remaining datasets, we adhere to the metrics specified in [Lee *et al.*, 2023b]. With the MM-IMDb dataset, the multi-label classification performance is assessed using F1-Macro. The classification accuracy is employed for the UPMC Food-101 dataset. Lastly, for Hateful Memes, the evaluation is based on the AU-ROC metric.

3.2 Baselines and Experimental Settings

Baselines: We incorporate several state-of-the-art approaches representing popular strategies into our comparative evaluation. Specifically, GMC [Poklukar *et al.*, 2022b] serves as a contrastive learning-based approach, ActionMAE [Woo *et al.*, 2023] represents a generation-based method, and we include a Transformer-based approach proposed in [Lee *et al.*, 2023b], referred to as Prompt-Trans for brevity. To ensure optimal reproducibility, we inherit the implementations of all baseline methods from their original code bases. Additionally, we implement unimodal frameworks (Unimodal) for each modality, trained in a supervised manner with available labels, to serve as our baseline comparison.

Implementation details: To ensure a fair comparison, we use similar encoder architectures for processing raw data modalities whenever possible. The unimodal baselines are designed with the same architectures as Robult, each with its own classifier. For Robult, positive samples for the soft P-U loss are determined after discretizing labels if needed. Specifically, in the cases of CMU-MOSI and CMU-MOSEI datasets, label information in the range of $[-3, 3]$ is quantified into 7 discrete categories $(-3, -2, \ldots, 3)$. Additionally, for the multi-label dataset MM-IMDb, two samples are considered positive if they share all the same labels. Regarding Prompt-Trans, we only report its results for three datasets involving two modalities (MM-IMDb dataset, UPMC Food-101 dataset, and Hateful Memes dataset), as the extension to multiple modalities cannot be directly inferred from the original work [Lee *et al.*, 2023b].

Experimental details: The primary focus of our performance reporting is on two extreme scenarios: semi-supervised settings with only 5% labeled data and scenarios where only a single modality is presented during evaluation. All reported results are averaged over 3 different random seeds. In the semi-supervised setup, the newly created labeled portion is ensured to maintain the correct label ratio as the original training sets. Additional experiments extending these two settings to more modalities and a higher percent of labeled data are detailed in Appendix D.1 and D.2 respectively. Specific details on implementation settings relating to each dataset are provided in Appendix - C.2.

	CMU-MOSI				CMU-MOSEI			
Metrics	Unimodal	GMC	ActionMAE	**Robult**	Unimodal	GMC	ActionMAE	**Robult**
Text Modality:								
MAE (↓)	1.41	1.407	1.476	**1.397**	0.81	0.815	1.115	**0.784**
Corr (↑)	0.137	0.14	0.066	**0.144**	0.383	0.346	0.136	**0.459**
F1 (↑)	0.551	0.559	0.535	**0.578**	0.717	0.716	0.614	**0.739**
Acc (↑)	0.553	0.562	0.47	**0.569**	0.712	0.708	0.603	**0.732**
Audio Modality:								
MAE (↓)	1.576	1.518	1.546	**1.415**	0.842	0.836	1.215	**0.825**
Corr (↑)	0.041	-0.065	0.046	**0.085**	0.111	0.193	0.101	**0.221**
F1 (↑)	0.512	0.457	0.508	**0.539**	0.618	0.642	0.634	**0.679**
Acc (↑)	0.496	0.46	0.467	**0.535**	0.599	0.63	0.543	**0.65**
Vision Modality:								
MAE (↓)	1.451	1.497	1.511	**1.425**	0.891	0.839	1.127	**0.826**
Corr (↑)	0.044	-0.07	-0.03	**0.086**	0.163	0.2	0.104	**0.201**
F1 (↑)	0.585	0.446	0.511	**0.593**	0.637	0.621	0.594	**0.647**
Acc (↑)	0.425	0.449	0.514	**0.522**	0.624	0.62	0.561	**0.632**
Full Modality:								
MAE (↓)	1.394	1.47	1.496	**1.392**	0.783	0.819	1.103	**0.779**
Corr (↑)	0.186	0.101	-0.092	**0.247**	0.364	0.328	0.337	**0.504**
F1 (↑)	0.597	0.497	0.553	**0.657**	0.73	0.693	0.694	**0.744**
Acc (↑)	0.594	0.498	0.477	**0.63**	0.729	0.688	0.643	**0.741**

Table 1. Results on CMU-MOSI, CMU-MOSEI.

3.3 Main Quantitative Results

All results are shown in tables with the best outcomes in **red** and the second-best in blue.

Sentiment Analysis: The results for CMU-MOSI and CMU-MOSEI datasets are summarized in Table 1. For both datasets, Robult significantly outperforms all the compared methods, suggesting its effectiveness and consistency in semi-supervised and missing modality scenarios. Regarding CMU-MOSI, due to its smaller scale compared to CMU-MOSEI, the labeled portions are also smaller. This condition poses a challenge for existing baselines that heavily rely on label information. In contrast, Robult effectively addresses this challenge, demonstrating the ability to extract meaningful representations even with limited labeled data. On CMU-MOSEI, Robult consistently produces superior representations, achieving the best performances across all recorded metrics. Notably, Robult improves the correlation (Corr) between the predicted sentiment levels and ground truth by up to 19.8%, outperforming the second-best method, which is the unimodal for textual data.

Classification tasks: In Table 2, empirical results for three classification tasks show that Robult consistently outperforms existing approaches and baselines in most cases, except for one scenario on the Hateful Memes dataset with the full modality available, where Robult achieves comparable performance with Prompt-Trans [Lee *et al.*, 2023b]. Notably, the Hateful Memes dataset includes samples with “benign confounders”, negatively impacting performance when models rely solely on single modalities [Kiela *et al.*, 2020]. Leveraging the soft Positive-Unlabelled loss, Robult effectively addresses and mitigates performance gaps with either single modality inputs or the full ones. In addition, we calculate F1 macro scores for all methods on these three datasets in the unimodal and multimodal cases. We further visualize a Critical Difference Diagram [Demšar, 2006] in Figure 5. This diagram visually represents the performance among different machine learning algorithms across various datasets by displaying the mean performance ranks, with lower being better,

	Unimodal	Prompt-Trans	GMC	ActionMAE	**Robult**
	MM-IMDb - F1 Macro (↑):				
Text	0.24	0.198	0.296	0.055	**0.321**
Image	0.207	0.148	0.291	0.039	**0.298**
Full	0.196	0.268	0.307	0.171	**0.332**
	UPMC Food-101 - Accuracy (↑):				
Text	0.321	0.151	0.395	0.196	**0.435**
Image	0.296	0.111	0.382	0.132	**0.415**
Full	0.138	0.432	0.41	0.358	**0.446**
	Hateful Memes - AUROC (↑):				
Text	0.584	0.511	0.617	0.528	**0.623**
Image	0.524	0.475	0.528	0.508	**0.596**
Full	0.618	**0.635**	0.616	0.542	0.632

Table 2. Results on MM-IMDb, UPMC Food-101, Hateful Memes.

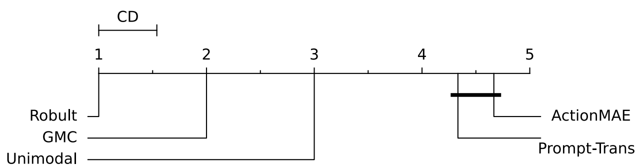


Figure 5. CD diagram showing the mean rank of each method on three datasets.

and connecting statistically indistinguishable groups (within 95% confidence level) with a thin horizontal bar, as per the Friedman hypothesis test. From the diagram, Robult exhibits a clear improvement gap compared to other state-of-the-art methods in average ranks, while ActionMAE and Prompt-Trans show no statistically significant difference in their performance.

3.4 Main ablation studies

Owing to space constraint, we provide here a key analysis of ablation study for different components of Robult. For a more comprehensive analysis, please refer to the additional experiments in Appendix D, which offer further insights into how architectural choices, Soft-PU loss, and weighting schemes influence Robult's performance quantitatively and qualitatively.

We evaluate the impact of each loss component on Robult's performance using Hateful Memes dataset, which mirror the semi-supervised and missing modalities conditions of our main experiments. This analysis involves testing variations of Robult with different ablations. (1) *Removal of* $\mathcal{L}_{sup}$ - this setting utilizes available label information only in $\mathcal{L}_{(u)lb}$, so Robult can only produce latent representations. An additional Logistic Regressor is trained with these representations as its input, and this pipeline's final scores are reported. (2) *Removal of* $\mathcal{L}_{rec}$ - this setting discards $\mathcal{L}_{rec}$, corresponding to our Objective 2.2. (3) *Removal of* $\mathcal{L}_{lb}$ - this setting makes the learning of Objective 2.1 rely only on $\mathcal{L}_{ulb}$. (4) *Removal of* $\mathcal{L}_{ulb}$ - this setting associates Objective 2.1 exclusively with $\mathcal{L}_{lb}$. Table 3 summarizes the results of this ablation experiment. Overall, any ablation negatively impacts the performance of Robult. In particular, the absence of $\mathcal{L}_{sup}$ significantly worsens the performance, as there is no loss guiding the learning of Robult's classifier, which is crucial for generating soft label information consumed by the soft Positive-Unlabeled loss $\mathcal{L}_{ulb}$. Consequently, this ablation adversely affects two loss components, explaining the poorest result among all variations. The removal of $\mathcal{L}_{rec}$ particularly harms the performance with unimodal inputs, aligning with the motivation for Objective 2.2, as the unique information U^i is no longer preserved. In two remaining cases, both ablations diminish Robult's overall performance, indicating their equal contribution to achieving Objective 2.1.

Metrics	GMC	**Robult**	Robult (1)	Robult (2)	Robult (3)	Robult (4)
	CMU-MOSI - Text Modality:					
MAE	1.407	**1.397**	1.589	1.511	1.443	1.429
Corr	0.14	**0.144**	0.047	0.101	0.051	0.123
F1	0.559	0.578	0.542	0.52	**0.593**	0.571
Acc	0.562	0.569	0.544	0.523	0.422	**0.573**
	CMU-MOSI - Audio Modality:					
MAE	1.518	**1.415**	1.586	1.561	1.494	1.495
Corr	-0.065	**0.085**	0.023	0.005	0.046	**0.085**
F1	0.457	**0.539**	0.526	0.499	0.51	0.517
Acc	0.46	**0.535**	0.518	0.502	0.442	0.509
	CMU-MOSI - Vision Modality:					
MAE	1.497	**1.425**	1.663	1.711	1.445	1.504
Corr	-0.07	**0.086**	0.025	-0.023	0.041	-0.066
F1	0.446	**0.593**	0.519	0.485	0.571	0.459
Acc	0.449	**0.522**	0.519	0.465	0.47	0.448
	CMU-MOSI - Full Modality:					
MAE	1.47	**1.392**	1.588	1.434	1.411	1.459
Corr	0.101	**0.247**	0.071	0.239	0.166	0.229
F1	0.497	**0.657**	0.524	0.567	0.549	0.6
Acc	0.498	**0.63**	0.523	0.566	0.552	0.601
	Hateful Memes - Text Modality:					
AUROC	0.617	**0.623**	0.528	0.59	0.605	0.576
Acc	0.581	**0.59**	0.535	0.556	0.571	0.562
	Hateful Memes - Image Modality:					
AUROC	0.528	**0.596**	0.518	0.582	0.588	0.566
Acc	0.551	**0.562**	0.524	0.551	0.526	0.539
	Hateful Memes - Full Modality:					
AUROC	0.616	0.632	0.538	0.618	**0.634**	0.582
Acc	0.532	**0.595**	0.542	0.55	0.554	0.552

Table 3. Ablation analysis on CMU-MOSI and Hateful Memes datasets for Robult.

4 Contributions & Limitations

Contributions: Our Robult pipeline leverages limited label data through a soft Positive-Unlabelled (PU) loss and latent reconstruction loss, enhancing modality interactions and preserving unimodal data integrity. It supports various modality types and quantities, scales linearly with modalities, and functions independently of specific architectures. This flexibility facilitates integration with existing DL frameworks, advancing multimodal learning in practical settings.

Limitations. Robult's design presumes that the proximity of positive couplets follows a Gaussian distribution, a method proven empirically but not theoretically. Future work should seek theoretical validation for this assumption. Moreover, with our setting, the potential of labeled data in scenarios with missing modalities in training remains untapped. Exploring these cases could further improve Robult's effectiveness in complex real-world applications.

Acknowledgements

The work of Duy A. Nguyen was supported in part by a PhD fellowship from the VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam.

References

[Aberdam *et al.*, 2022] Aviad Aberdam, Roy Ganz, Shai Mazor, and Ron Litman. Multimodal semi-supervised learning for text recognition. *arXiv preprint arXiv:2205.03873*, 2022.

[Acosta *et al.*, 2022] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

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

[Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NeurIPS*, 29, 2016.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2021] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7016–7025, 2021.

[Chen *et al.*, 2023] Mengxi Chen, Linyu Xing, Yu Wang, and Ya Zhang. Enhanced multimodal representation learning with cross-modal kd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11766–11775, 2023.

[Daunhawer *et al.*, 2023] Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. *arXiv preprint arXiv:2303.09166*, 2023.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

[Feichtenhofer *et al.*, 2022] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *NeurIPS*, 35:35946–35958, 2022.

[Huang *et al.*, 2021] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *NeurIPS*, 34:10944–10956, 2021.

[Kiela *et al.*, 2020] 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. *NeurIPS*, 33:2611–2624, 2020.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594. PMLR, 2021.

[Kingma and Ba, 2015] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2014.

[Lee and Van der Schaar, 2021] Changhee Lee and Mihaela Van der Schaar. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics*, pages 1513–1521. PMLR, 2021.

[Lee *et al.*, 2023a] Seungyeol Lee, Taeho Kim, and Jae-Pil Heo. Cross-loss pseudo labeling for semi-supervised segmentation. *IEEE Access*, 2023.

[Lee *et al.*, 2023b] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023.

[Li *et al.*, 2021] Shuang Li, Wenxuan Ma, Jinming Zhang, Chi Harold Liu, Jian Liang, and Guoren Wang. Meta-reweighted regularization for unsupervised domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2781–2795, 2021.

[Liang *et al.*, 2023] Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. *arXiv preprint arXiv:2306.04539*, 2023.

[Ma *et al.*, 2021] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 3 in 35, pages 2302–2310, 2021.

[Ma *et al.*, 2022] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

[Peng *et al.*, 2005] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022.

[Poklukar *et al.*, 2022a] Petra Poklukar, Vladislav Polianskii, Anastasiia Varava, Florian T Pokorny, and Danica Kragic. Delaunay component analysis for evaluation of data representations. In *International Conference on Learning Representations; Apr 25th through Fri the 29th, 2022 (online).*, 2022.

[Poklukar *et al.*, 2022b] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *ICML*, pages 17782–17800. PMLR, 2022.

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

[Su *et al.*, 2021] Mingyue Su, Guanghua Gu, Xianlong Ren, Hao Fu, and Yao Zhao. Semi-supervised knowledge distillation for cross-modal hashing. *IEEE Transactions on Multimedia*, 2021.

[Tsai *et al.*, 2018] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[Tsai *et al.*, 2019] 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. NIH Public Access, 2019.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wang and Isola, 2020] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 13–18 Jul 2020.

[Wang *et al.*, 2015] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015.

[Wang *et al.*, 2020] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020.

[Wang *et al.*, 2023] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *CVPR*, pages 22025–22034, 2023.

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

[Woo *et al.*, 2023] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 3 in 37, pages 2776–2784, 2023.

[Wu and Goodman, 2018] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *NeurIPS*, 31, 2018.

[Xu *et al.*, 2024] Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Hengtao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. *NeurIPS*, 36, 2024.

[Zadeh *et al.*, 2016] 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.

[Zadeh *et al.*, 2018a] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

[Zhang *et al.*, 2022a] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669, 2022.

[Zhang *et al.*, 2022b] Xinyu Zhang, Zhiwei Li, Yan Gong, Dafeng Jin, Jun Li, Li Wang, Yanzhang Zhu, and Huaping Liu. Openmpd: An open multimodal perception dataset for autonomous driving. *IEEE Transactions on Vehicular Technology*, 71(3):2437–2447, 2022.

[Zhang *et al.*, 2023] Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, and Zhenghua Xu. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 83:102656, 2023.