RESEARCH-ARTICLE

PgM: Partitioner Guided Modal Learning Framework

GUIMIN HU, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

YI XIN, Nanjing University, Nanjing, Jiangsu, China

LIJIE HU, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, Abu Dhabi, United Arab Emirates

ZHIHONG ZHU, Tencent, Shenzhen, Guangdong, China

HASTI SEIFI, Arizona State University, Tempe, AZ, United States

Open Access Support provided by:

University of Copenhagen

Mohamed Bin Zayed University of Artificial Intelligence

Tencent

Nanjing University

Arizona State University

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

PgM: Partitioner Guided Modal Learning Framework

Guimin Hu
Guangdong University of Technology
Guangzhou, China
University of Copenhagen
Copenhagen, Denmark
rice.hu.x@gmail.com

Yi Xin
Nanjing University
Nanjing, China
xin.yi@smail.nju.edu.cn

Lijie Hu
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE
lijie.hu@mb.zuai.ac.ae

Zhihong Zhu
Tencent
Shenzhen, China
profzhu@tencent.com

Hasti Seifi
Arizona State University
Tempe, USA
hasti.seifi@asu.edu

Abstract

Multimodal learning benefits from multiple modal information, and each learned modal representations can be divided into *uni-modal* that can be learned from uni-modal training and *paired-modal* features that can be learned from cross-modal interaction. Building on this perspective, we propose a partitioner-guided modal learning framework, PgM, which consists of the modal partitioner, uni-modal learner, paired-modal learner, and uni-paired modal decoder. Modal partitioner segments the learned modal representation into uni-modal and paired-modal features. Modal learner incorporates two dedicated components for uni-modal and paired-modal learning. Uni-paired modal decoder reconstructs modal representation based on uni-modal and paired-modal features. **PgM offers three key benefits**: 1) thorough learning of uni-modal and paired-modal features, 2) flexible distribution adjustment for uni-modal and paired-modal representations to suit diverse downstream tasks, and 3) different learning rates across modalities and partitions. Extensive experiments demonstrate the effectiveness of PgM across **four multimodal tasks** and further highlight its transferability to existing models. Additionally, we visualize the distribution of uni-modal and paired-modal features across modalities and tasks, offering insights into their respective contributions.

CCS Concepts

• **Computing methodologies** → *Artificial intelligence*; *Artificial intelligence*.

Keywords

Modal learning, Modal laziness, Multimodal learning framework, Uni-modal, Paired-modal

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754788

ACM Reference Format:
Guimin Hu, Yi Xin, Lijie Hu, Zhihong Zhu, and Hasti Seifi. 2025. PgM: Partitioner Guided Modal Learning Framework. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754788

1 Introduction

Multimodal data is widely used to enhance machine learning systems, which integrates these diverse modalities to improve decision-making accuracy [9–11, 37, 38]. With the inclusion of additional information, multimodal networks are expected to match or surpass their uni-modal counterparts, as they leverage richer information from multiple input modalities [12, 32].

However, multimodal networks are often observed to underperform uni-modal networks. This observation is consistent across various modality combinations, tasks, and benchmarks [12, 13, 18, 30]. Liang et al. [20] analyzes multimodal model behavior by studying the uni-modal importance, cross-modal interactions and so on, and proposes that the gradients of different modalities should be further adjusted during training. Du et al. [4] also notice the phenomenon of modality laziness, which causes insufficient modal learning of uni-modal feature. Wang et al. [30] identifies multimodal network are ofen prone to overfiting due to their increased capacity and different modalities overfit and generalize at different rates.

Previous multimodal learning studies summarize three major unresolved challenges. **First**, different types of features in multimodal data are learned at varying rates. Furthermore, The uni-modal and paired-modal representations required for each modality vary across different tasks [4, 30]. **Second**, achieving an effective multimodal fusion representation requires both robust learning of uni-modal and paired-modal features and dynamic adjustment of their distributions throughout the representation learning process [12]. **Third**, modality laziness causes insufficient modal learning since multimodal networks are often prone to overfitting and generalizing at different rates [4].

Building on [14] and considering how learned representations are formed from multimodal data in supervised learning, we introduce a modal partitioner (Figure 1), where uni-modal and paired-modal partitions can overlap and are not mutually exclusive. From this perspective, we propose a **P**artitioner-**G**uided **M**odal learning

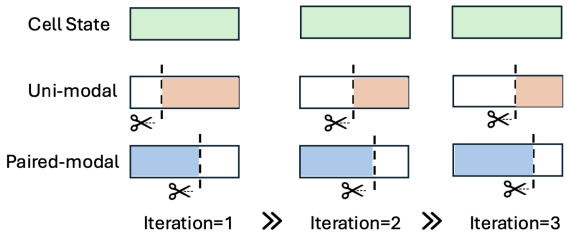

Figure 1: Illustration of modal partitioner. Modal partitioner adjusts the distribution of uni-modal and paired-modal representations across multiple iterations.

framework, **PgM**, which comprises a modal partitioner for segmentation, two modal learner for modal partition learning, and a modal decoder for modality reconstruction. The proposed modal partitioner is inspired by the gate mechanism in [26], widely applied in various tasks [2, 33]. Using a cumulative softmax-based binary gate, it splits the cell state into two segments (0 and 1), allowing different update rules to distinguish uni-modal from paired-modal information. However, since the segmentation is not based on semantic roles, the gating mechanism does not align with them. The modal partitioner first segments the learned modal representation into uni-modal and paired-modal partitions. Next, the modal learner includes two dedicated components for learning uni-modal and paired-modal representations, while the modal decoder reconstructs modality information between these features to effectively capture their characteristics. Finally, PgM is integrated with downstream tasks in an end-to-end architecture, enabling joint training. **PgM offers three key advantages**: 1) it enables efficient learning of both uni-modal and paired-modal features, 2) it allows for flexible adaptation of feature distributions across different downstream tasks while dynamically adjusting uni-modal and paired-modal contributions, and 3) it tailors learning rates across modalities and partitions to counteract modality laziness. To sum up, our contributions are three-folds:

1. We propose PgM, a partitioner-guided modal learning framework, consists of the modal partitioner, two modal learners (e.g., uni-modal and paired-modal learners), and a uni-paired modal decoder for effective modal partition representation learning.
2. PgM can be integrated into existing models for multimodal downstream tasks, demonstrating its transferability and feasibility.
3. Experimental results demonstrate the effectiveness of PgM across four multimodal tasks and its transferability to existing models, providing deeper insights into multimodal learning.

2 Related Work

2.1 Multimodal Networks

Multimodal networks process multimodal signals as input and integrate the information to support decision-making in downstream tasks. These networks either use one modality as input to predict another (e.g., Visual-Q&A [6]) or leverage cross-modality correspondences for self-supervised learning (e.g., image-audio correspondence [1]). Existing multimodal networks can be broadly classified into three categories: Transformer-based architectures, disentangle-based methods and generation-based structures. First is Transformer-based architectures. For instance, Tsai et al. [28] introduces the Multimodal Transformer (MulT), which addresses multimodal fusion in an end-to-end manner. Second is disentangle-based methods. These approaches separate modal representations into modality-invariant and modality-specific feature spaces. For example, Zhang et al. [36] introduces an adversarial multimodal refinement module to explore shared characteristics across modalities while enhancing each modality's uniqueness. Third is generation-based structures that utilize translation modules to generate modal representations from another modality. More recently, models like BLIP [16] and BLIP-2 [15] have been proposed for vision-language tasks, excelling in both understanding and generation.

2.2 Multimodal Learning

Multimodal learning involves integrating information from different modalities, which can be categorized into three types: early fusion, late fusion, and hybrid fusion. Early fusion aggregates the raw features from different modalities into a joint representation before modal encoder [22]. Late fusion combines the decisions from different classifiers into one final decision Grover et al. [7]. Compared with early fusion and late fusion, hybrid fusion is a multimodal learning approach that combines elements of both early fusion and late fusion, aiming to leverage the strengths of both methods while mitigating their weaknesses [10, 11, 37]. Recently, Wang et al. [30] observes that the best-performing uni-modal network often outperforms its multimodal counterpart. This phenomenon is consistent across various modality combinations, tasks, and benchmarks in video classification. Similarly, Du et al. [4] highlights the issue of modality laziness, where modal representation fails to learn effectively. Huang et al. [13] further demonstrates that in multimodal late-fusion networks with (smoothed) ReLU activation trained via gradient descent, different modalities compete, leading the encoder networks to focus on only a subset of modalities.

The prior work underscore that different modalities exhibit varying learning dynamics during training, influencing final model performance. Some studies [20] analyze multimodal model behavior by examining uni-modal importance, cross-modal interactions, and other factors, arguing that modality-specific gradient adjustments are necessary. Additionally, uni-modal features in multimodal fusion representations may be insufficient due to significant redundancy—often repetitive or uninformative—in the final fused representation [12].

3 Methodology

3.1 Architecture of PgM

In this section, we will introduce the architecture of PgM, and its architecture is illustrated in Figure 2(b). We first provide the components of PgM and then show the specific architecture.

Modal Encoder: We adopt T5 [24] as text encoder, AST [5] as audio encoder and ViT [3] as image encoder. Each modal encoder take raw modal information as input and outputs modal representation

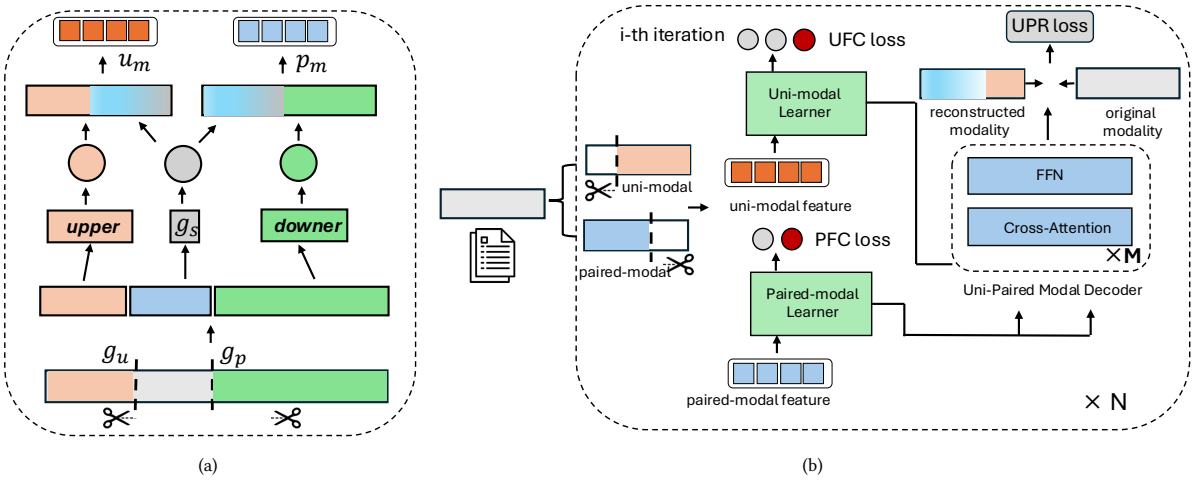

Figure 2: Overview of PgM: (a) The partitioner performs segmentation by processing modal representations and outputting uni-modal u_m and paired-modal p_m features. (b)The PgM architecture consists of a modal partitioner, a uni-modal learner, a paired-modal learner, and a uni-paired modal decoder. These components are jointly trained using three pre-training objectives: uni-modal feature classification, paired-modal feature classification, and uni-paired modal reconstruction.

as shape (B,S,D), where B denotes batch size, S denotes sequence length, and D denotes the dimension of modal representation.

Modal Partitioner: The modal partitioner takes each individual modal representation as input and outputs uni-modal and paired-modal gates with a shape of (B,S,D). Furthermore, we represent uni-modal and paired-modal gates as a vector, where each element belongs to $\{0, 1\}$. These gates are then used to derive padding masks for the uni-modal and paired-modal learners, respectively.

Uni-Modal Learner: The uni-modal learner learns uni-modal features by masking paired-modal features, which are indicated by $-\infty$ in the padding mask. The padding mask has a shape of (B,S,D) corresponding to the uni-modal learner.

Paired-Modal Learner: The paired-modal learner learns paired-modal features by masking uni-modal features, which are indicated by $-\infty$ in the padding mask of the paired-modal learner. The padding mask has a shape of (B,S,D) corresponding to the paired-modal representation.

Uni-Paired Decoder: The uni-paired modal decoder takes the concatenation of the uni-modal and paired-modal representations as input, then passes it through the subsequent FFN layer to reconstruct the original modal representation.

Padding Mask We derive padding masks for uni-modal and paired-modal learners from $\hat{\mathbf{g}}_u^{(i)}$ and $\hat{\mathbf{g}}_p^{(i)}$, respectively. For instance, given $\hat{\mathbf{g}}_u^{(i)} \in \mathcal{R}^{1\times D}$ with $\hat{\mathbf{g}}_u^{(i)} = [1, 1, 1, 1, 0, 0]$, the first four positions indicate the features the uni-modal learner should focus on. Accordingly, its padding mask $\mathbf{M}^u \in \mathcal{R}^{S\times D}$ is generated through a broadcasting operation on $\hat{\mathbf{g}}_u^{(i)}$, where S and D denote the sequence length and representation dimension, respectively. Similarly, we also can obtain padding mask $\mathbf{M}^p \in \mathcal{R}^{S\times D}$ for paired-modal learner.

3.2 Modal Partitioner

Given a sample containing multiple modalities $\{m^1, m^2, \cdots, m^N\}$ (e.g., text modalities), we set N individual modal encoders to map raw modal signal into a d-dimensional representation vector. After modal encoder, we derive the modality representation set $\{\mathbf{I}_{m^1}, \cdots, \mathbf{I}_{m^N}\}$ for N modalities. Figure 2(a) provides segmentation process by the partitioner.

The modal partitioner processes modal representations by partitioning neurons into two groups: uni-modal and paired-modal features. The uni-modal features store information for uni-modal training, while the paired-modal features capture information for cross-modal interactions [4]. We assign a modal partitioner to each modality, dividing each modal representation into two partitions. Given a modality representation $\mathbf{I}_m$, the segmentation point between uni-modal and paired-modal features is determined using the cumulative softmax activation function, defined as $\text{cumsoftmax}(\cdot) = \text{cumsum}(\text{softmax}(\cdot))$[1], which is referred to as master gates in [26]. The cumsoftmax function dynamically adjusts the cut-off points based on the input, enabling flexible partitioning and approximating a binary gating mechanism of the form $(0, \cdots, 1, \cdots 1)$. In the i-th iteration, the partitioner gates for uni-modal and paired-modal features are defined as follows:

$$\begin{aligned} \mathbf{g}_u^{(i)} &= 1 - \text{cumsoftmax}(\mathbf{u}_m) \\ \mathbf{g}_p^{(i)} &= \text{cumsoftmax}(\mathbf{p}_m) \end{aligned} \tag{1}$$

The intuition behind Equation (1) is to identify two cut-off points, illustrated as scissors in Figure 1, that naturally partition the set of neurons into two distinct segments. Initially, in the first iteration, $\mathbf{u}_m = \mathbf{p}_m = \mathbf{I}_m$, $m \in \{m^1, m^2, \cdots, m^N\}$ denotes specific modality. $\mathbf{g}_u^{(1)}$ and $\mathbf{g}_p^{(1)}$ denote the uni-modal and paired-modal gates in

[1] $\text{cumsoftmax}(x_i) = \sum_{j=1}^{i} \frac{e^{x_j}}{\sum_{k=1}^{n} e^{x_k}}$.

the first iteration, respectively. In the i-th iteration,$\mathbf{g}_u^{(i)} \in \mathcal{R}^{1\times D}$ follows the form $[u_1, \cdots, u_r, 0, \cdot, 0]$, and $\mathbf{g}_p^{(i)} \in \mathcal{R}^{1\times D}$ takes the form $[0, \cdot, 0, p_1, \cdots, p_l]$. These gates control the uni-modal and paired-modal features in the i-th iteration, where $u_i \rightarrow 1$ and $p_i \rightarrow 1$.

$$\begin{aligned} \mathbf{g}_s^{(i)} &= \mathbf{g}_u^{(i)} \circ \mathbf{g}_p^{(i)} \\ \mathbf{upper}_m^{(i)} &= \mathbf{g}_u^{(i)} - \mathbf{g}_s^{(i)} \\ \mathbf{downer}_m^{(i)} &= \mathbf{g}_p^{(i)} - \mathbf{g}_s^{(i)} \end{aligned} \tag{2}$$

where $\circ$ denotes element-wise multiplication, $\mathbf{g}_s^{(i)}$ represents the gate overlap between uni-modal and paired-modal features. $\mathbf{upper}_m^{(i)}$ refers to the upper part of the cell neurons representing uni-modal features, while $\mathbf{downer}_m^{(i)}$ refers to the lower part of the cell neurons representing paired-modal features. $\mathbf{I}_s^m$ denotes the shared modal representation between uni-modal and paired-modal. Subsequently, we aggregate partition information from both target cells to obtain the updated uni-modal and paired-modal representations:

$$\begin{aligned} \mathbf{I}_s^m &= \mathbf{g}_s^{(i)} \circ \mathbf{I}_m \\ \mathbf{u}_m^{(i+1)} &= \mathbf{upper}_c \circ \mathbf{u}_m^{(i)} + \mathbf{I}_s^m \\ \mathbf{p}_m^{(i+1)} &= \mathbf{downer}_c \circ \mathbf{p}_m^{(i)} + \mathbf{I}_s^m \end{aligned} \tag{3}$$

where $\{\mathbf{u}_m^{(i+1)}, \mathbf{p}_m^{(i+1)}\}$ are updated uni-modal and paired-modal features after the i-th iteration, respectively. Thus, after N iterations, the modal partitioner divides each learned modal representation into the uni-modal partition $\mathbf{u}_m^{(N)}$ and the paired-modal partition $\mathbf{p}_m^{(N)}$.

3.3 Modal Learner and Decoder

The modal learner consists of (1) uni-modal learner and (2) paired-modal learner, and each learner contains multiple stacked Transformer blocks. To enhance uni-modal and paired-modal learning, we apply padding masks to specific neurons within the multimodal representation, ensuring that the uni-modal learner and paired-modal learner focus exclusively on uni-modal and paired-modal features, respectively. In the i-th iteration, the padding masks are derived from $\mathbf{g}_u^{(i)}$ and $\mathbf{g}_p^{(i)}$ to tailor the attention masks for the two learners. For example, $\mathbf{g}_u^{(i)} = [u_1, \cdots, u_m, 0, \cdots, 0]$, $\mathbf{g}_p^{(i)} = [0, \cdot, 0, p_1, \cdots, p_m]$, and $u_i \rightarrow 1, p_i \rightarrow 1$, thus we can approximately formalize $\mathbf{g}_u^{(i)}$ and $\mathbf{g}_p^{(i)}$ as $[1, \cdots, 1, 0, \cdots, 0]$ and $[0, \cdots, 0, 1, \cdots, 1]$, denoted by $\hat{\mathbf{g}}_u^{(i)}$ and $\hat{\mathbf{g}}_p^{(i)}$, respectively:

$$\mathbf{M}_z^{(i)} = (\mathbf{1.0} - \hat{\mathbf{g}}_z^{(i)}) \cdot C, \quad z \in \{u, p\} \tag{4}$$

$$\mathbf{A}_z^{(i)} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{M}_z^{(i)}) \tag{5}$$

where constant C=-10000.0 matches the data's dimensionality and serves as a tuning parameter in the model, used to amplify certain operations. $\mathbf{M}_z^{(i)}$ denotes padding mask, $\{\mathbf{A}_u^{(i)}, \mathbf{A}_p^{(i)}\}$ represents the attention weight matrix applied to the Transformer [29] modules for the uni-modal and paired-modal learners, respectively.

Uni-modal and Paired-modal Learners. uni-modal learner focuses on intra-uni-modal feature learning, and paired-modal learner attends on intra-paired-modal learning. Both learners take Transformer as backbone, which takes modal representation as query, key and value. Noted that uni-modal and paired-modal learners takes $\mathbf{M}^u \in \mathcal{R}^{S\times D}$ and $\mathbf{M}^p \in \mathcal{R}^{S\times D}$ as padding masks respectively, ensuring the learners only focus on uni-modal and paired-modal features. Here, S and D denote the sequence length and representation dimension, respectively. The padding masks for uni-modal features (i.e., $\mathbf{M}^u$) and paired-modal features (i.e., $\mathbf{M}^p$) are defined as follows, respectively:

$$M_{*j}^u = \begin{cases} 0, & j <= u \\ -\infty, & j > u \end{cases}, \qquad M_{*j}^p = \begin{cases} 0, & j >= p \\ -\infty, & j < p \end{cases} \tag{6}$$

The illustrations of $\mathbf{M}^u$ and $\mathbf{M}^p$ are give by:

$$\mathbf{M}^u = \begin{bmatrix} 0 & \cdots & 0 & \downarrow u & -\infty & \cdots & -\infty \\ 0 & \cdots & 0 & & -\infty & \cdots & -\infty \\ 0 & \cdots & 0 & & -\infty & \cdots & -\infty \\ 0 & \cdots & 0 & & -\infty & \cdots & -\infty \end{bmatrix}$$

$$\mathbf{M}^p = \begin{bmatrix} -\infty & \cdots & -\infty & \downarrow p & 0 & \cdots & 0 \\ -\infty & \cdots & -\infty & & 0 & \cdots & 0 \\ -\infty & \cdots & -\infty & & 0 & \cdots & 0 \\ -\infty & \cdots & -\infty & & 0 & \cdots & 0 \end{bmatrix}$$

where u and p denote the upper and lower segmentation points of the uni-modal and paired-modal partitions, respectively. Both are derived from $\hat{\mathbf{g}}_u^{(i)}$ and $\hat{\mathbf{g}}_p^{(i)}$.

Uni-paired Modal Decoder. The uni-paired modal decoder takes the concatenated uni-modal and paired-modal features to reconstruct the modal representation.

$$\hat{\mathbf{I}}_m^{(i)} = \text{Decoder}([\mathbf{u}_m^{(i)}, \mathbf{p}_m^{(i)}]) \tag{7}$$

where $[,]$ denotes concatenation operation, $\hat{\mathbf{I}}_m^{(i)}$ denotes the reconstructed modal representation produced by uni-paired modal decoder.

3.4 Pre-training Objectives

The PgM architecture is illustrated in Figure 2(b). The framework is trained with three objectives: **Uni-modal Feature Classification**, which enhances uni-modal feature learning; **Paired-modal Feature Classification**, which focuses on paired-modal feature learning; and **Uni-paired Modal Reconstruction**, which reconstructs modalities by leveraging both uni-modal and paired-modal features within the same modality.

Uni-modal Feature Classification. aims to learn better uni-modal representations before performing modal fusion. We feed uni-modal feature of each iteration into classifier to estimate which modality the representation comes from, which ensures the uni-modal features store the essential information for discriminability of modality. For a sample with two modalities (e.g., text and vision),

the ground truth modality labels of text and vision modalities are denoted as O^t and O^v, respectively:

$$\mathbf{O}^t = \begin{bmatrix} 1 & 0 \\ \cdot & \cdot \\ 1 & 0 \end{bmatrix}, \mathbf{O}^v = \begin{bmatrix} 0 & 1 \\ \cdot & \cdot \\ 0 & 1 \end{bmatrix} \tag{8}$$

Similarly, we train an N-class classifier to determine the modality source of the uni-modal feature for a sample with N modalities. We denote the uni-modal feature classification loss $\mathcal{L}^{UFC}$ as the main supervision of uni-modal learner.

Paired-modal Feature Classification. To ensure the paired-modal feature remains distinct from the uni-modal feature, we input the paired-modal feature of each iteration into a classifier designed to differentiate between uni-modal and paired-modal representations. We formalize this process as a binary classification task, where the ground truth modality label is defined as:

$$\mathbf{O}^u = \begin{bmatrix} 1 & 0 \\ \cdot & \cdot \\ 1 & 0 \end{bmatrix}, \mathbf{O}^p = \begin{bmatrix} 0 & 1 \\ \cdot & \cdot \\ 0 & 1 \end{bmatrix} \tag{9}$$

where $\mathbf{O}^u$ and $\mathbf{O}^p$ denote the ground labels of uni-modal and paired-modal representation, respectively. We denote the paired-modal feature classification loss $\mathcal{L}^{PFC}$ as the main supervision of paired-modal learner.

Uni-Paired Modal Reconstruction. Each modal representation is decomposed into uni-modal and paired-modal components. The uni-paired modal decoder applies a reconstruction loss to ensure that the reconstructed modal representation captures both uni-modal and paired-modal features simultaneously, effectively reconstructing the original modal representation.

$$\mathcal{L}^{UPR} = \frac{1}{|S|}\left(\sum_{m\in\{\mathbf{m}^1,\cdots,\mathbf{m}^N\}} \frac{\|\mathbf{I}_m - \hat{\mathbf{I}}_m\|_2^2}{d_h}\right) \tag{10}$$

where $|S|$ represents the number of samples in the training set. We define the uni-paired modal reconstruction loss $\mathcal{L}^{UPR}$ as the primary supervisory signal for the uni-paired decoder.

3.5 Multimodal Downstream Tasks

During training, we integrate multimodal supervised downstream tasks and the partitioner-guided modal learning framework (PgM) for joint training and evaluate the performance on multimodal tasks. The training process comprises two stages: (1) pre-training the partitioner-guided modal learning framework (PgM) with the pre-training objective $\mathcal{L}^P$, and (2) fine-tuning PgM using both the downstream task loss and the pre-training objective loss $\mathcal{L}^D$:

$$\mathcal{L}^P = \sum_i^N (\mathcal{L}^{UFC} + \mathcal{L}^{PFC} + \mathcal{L}^{UPR}) \tag{11}$$

$$\mathcal{L}^D = \alpha\mathcal{L}^P + \beta\mathcal{L}^T \tag{12}$$

where $\mathcal{L}^T$ represents the training loss for downstream task T, $\{\alpha, \beta\}$ are the hyperparameters of PgM, i represents the i-th iteration and N is the iteration number. Modal partitioner adjusts the distribution of uni-modal and paired-modal representations across multiple iterations. After applying PgM, we obtain uni-modal and paired-modal features for each modality, enabling flexible distribution adjustment to better suit the given downstream task.

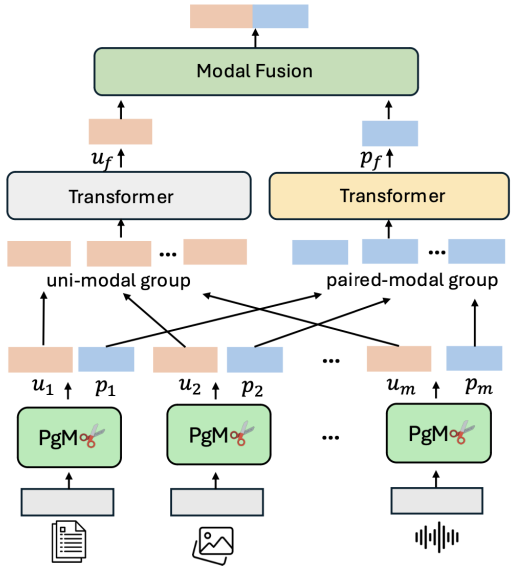

Figure 3: An overview of the training framework for multi-modal downstream tasks, which comprises PgM, modality-specific learning, and modality fusion components.

Detail Architecture. Figure 3 shows the architecture for jointly training PgM and the downstream task. First, PgM allows us to extract learned uni-modal and paired-modal features for each modality. These features are then grouped into two categories: uni-modal and paired-modal features across all modalities. For each sample, we independently concatenate all uni-modal features and all paired-modal features from different modalities, and pass the resulting representations through two separate feedforward (FFN) layers. Next, the concatenated uni-modal and paired-modal representations are fed into two distinct Transformer modules, where each serves as the query, key, and value. This step produces the final uni-modal and paired-modal representations, which integrate information from all modalities. For a sample with modalities $\{m_1, m_2\}$, the following modules in downstream task training are as follows:

$$\begin{aligned} \hat{\mathbf{p}} &= \text{FFN}([\mathbf{p}_{m_1}, \mathbf{p}_{m_1}]) \\ \hat{\mathbf{u}} &= \text{FFN}([\mathbf{u}_{m_1}, \mathbf{u}_{m_1}]) \\ \mathbf{p}_f &= \text{Transformer}(\hat{\mathbf{p}}, \hat{\mathbf{p}}, \hat{\mathbf{p}})) \\ \mathbf{u}_f &= \text{Transformer}(\hat{\mathbf{u}}, \hat{\mathbf{u}}, \hat{\mathbf{u}})) \end{aligned} \tag{13}$$

Here, $\{\mathbf{p}_{m_1}, \mathbf{p}_{m_2}\}$ represent the paired-modal features from modalities $\{m_1, m_2\}$, while $\{\mathbf{u}_{m_1}, \mathbf{u}_{m_2}\}$ represent the uni-modal features from modalities $\{m_1, m_2\}$. $\mathbf{u}_f$ and $\mathbf{p}_f$ denote the uni-modal and paired-modal representations, respectively, formed by combining multiple paired-modal features from different modalities. We pass

# Task	# Dataset	# Total Instances	# Number of Annotations	# Modality
Multimodal Sentiment Analysis	MOSI	2199	3	A+V+T
Multimodal Emotion Recognition	MELD	9989	6	A+V+T
Cross-modal Retrieval	Wikipedia	2866	10	V+T
Image-text Classification	UMPC Food 101	90686	101	V+T

Table 1: Information about the datasets used in four tasks. A, V, T denote Audio, Vision and Text modality.

the combined representations of uni-modal and paired-modal representation through two FFN layers:

$$\begin{aligned} \hat{\mathbf{I}} &= \text{FFN}([\mathbf{u}_f, \mathbf{p}_f]) \\ \hat{\mathbf{y}} &= \text{Prediction}(\hat{\mathbf{I}}) \end{aligned} \tag{14}$$

where $\hat{\mathbf{y}}$ denotes prediction results on categories. $\hat{\mathbf{I}}$ denotes the final multimodal fusion representation, where we use concatenation as the modal fusion method. The goal of the modal fusion module is to combine the uni-modal and paired-modal representations into a single vector. This can be achieved through fusion methods such as concatenation, addition, or more complex structures.

4 Experiments

4.1 Tasks and Datasets

In this work, we focus on the following tasks: **Multimodal Sentiment Analysis (MSA)** [10, 21] aims to predict the sentiment polarity by leveraging three types of modalities: audio, vision and text. **Multimodal Emotion Recognition in Conversation (MERC)** aims to predict predefined emotion categories (e.g., joy and sadness) using text, audio, and visual modalities. **Cross-modal Retrieval (CR)** [17, 27] is the process of finding the relevant items in one modality based on the query in another modality. **Image-text Classification (ITC)** [19] involves using both visual and textual information to classify the given information into 101 categories. We evaluate our proposed model on MSA, MERC, CR and ITC using MOSI [35], MELD [23], Wikipedia [31] and UMPC Food 101 [25] datasets, respectively. The details are shown in Table 1.

4.2 Experimental Setting and Evaluation Metrics

We use pre-trained T5-Base [2] as text encoder, ViT[3] as visual encoder, AST[4] as audio encoder. We set the learning rates to 3e-4 for the overall model and 1e-4 for the modal learners (i.e., uni-modal and paired-modal learners). The learning rate for the uni-paired decoder is set separately to 1e-3. In Equation (12), the hyperparameters are defined as $\alpha = 0.5$, $\beta = 1$ and the iteration number $N = 3$. The pre-training epoch for PgM is set to 20 (i.e., N^{1st}=20), and the jointly training epoch for downstream task and PgM is set to 50 (i.e.,N^{2st}=50). We adopt the Transformer architecture as the core backbone for uni-modal and paired modal learners, guided by the partitioner. For decoder, we adapt Cross-Attention layer for reconstruction. The dimensions of learned modal representation for each modality and the fused representation are set to 768. Following the previous work [8, 10], we adopt accuracy (ACC) and weighted F1 score (Weighted F1) to evaluate the performance on MELD and UPMC Food 101 datasets, and we adopt accuracy (ACC-2) and F1 as the metric of PgM on MOSI dataset. For Wikipedia dataset, we use the mean average precision (MAP) and Precision@10 (the average precision of the first 10 retrieved items) as the evaluation metric of cross-modal retrieval task.

[2]https://github.com/huggingface/transformers/tree/main/src/transformers/models/t5
[3]https://huggingface.co/openai/clip-vit-base-patch32
[4]https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer

4.3 Baselines

Concatenation concatenates feature vectors from different modalities into one feature vector. **Add** adds feature vectors from different modalities into one feature vector. **Element-wise Maximum** selects the element-wise maximum feature from different modalities. **Linear Fusion** applies linear combination of features from different modalities. **MLP** applies multilayer perceptron layer to fuse features from different modalities. **Uni-Modal Training** applies single modality for downstream task learning. Additionally, we further apply the proposed model fusion to the prior models of MSA and MERC tasks. The details of the prior models are as follows: **Self-MM** [34] leverages uni-modal representations through multi-task learning to address the multimodal sentiment analysis task. **MMIM** [8] hierarchically maximizes mutual information to tackle the multimodal sentiment analysis task. **UniMSE** [10] introduces a unified sharing framework that bridges multimodal sentiment analysis and multimodal emotion recognition to enhance model performance. **UniMEEC** [11] explores the complementary influence of emotion causes on multimodal emotion recognition.

4.4 Main Results

We present the experimental results of PgM alongside model learning baselines, as shown in Table 2. First, we conduct experiments to single-modality scenarios using the proposed framework and compare their performance. The results show that all multimodal fusion methods outperform their single-modal counterparts, emphasizing the importance of integrating complementary information from different modalities.

Next, we compare the performance between PgM and multimodal learning method (e.g., concatenation and add) across four multimodal tasks. Our proposed multimodal learning framework consistently surpasses all baseline multimodal learning methods across various tasks, highlighting its strong generalization and learning capabilities. For instance, in multimodal sentiment analysis, PgM outperforms baselines such as Concatenation and Add by at least 15%-18% points. In the multimodal emotion recognition task, PgM achieves an accuracy (ACC) of 66.69 and a weighted

#Model	#Multimodal Sentiment Analysis		#Multimodal Emotion Recognition		#Cross-modal Retrieval		#Image-text Classification	
	MOSI		MELD		Wikipedia		UMPC Food 101	
	ACC-2	**F1**	**ACC**	**Weighted F1**	**MAP**	**Precision@10**	**ACC**	**Weighted F1**
Single-Modal (Audio)	48.58/50.36	44.31/46.47	51.44	52.72	-	-	-	-
Single-Modal (Visual)	65.86/69.72	65.97/69.94	51.75	54.06	-	-	71.58	72.04
Single-Modal (Text)	64.33/66.16	63.20/65.32	56.21	56.38	-	-	69.92	69.89
Concatenation	69.15/70.42	67.26/69.17	58.64	58.16	62.84	56.29	82.43	80.49
Add	67.61/67.61	66.13/66.13	53.58	54.67	61.98	55.05	81.67	81.66
Element-wise Maximum	68.16/69.50	65.16/66.87	51.59	52.86	61.99	56.96	80.11	80.09
Linear-Fusion	66.52/67.81	64.58/65.36	53.72	54.89	61.54	54.14	77.10	77.05
MLP	66.30/67.82	65.41/66.57	54.16	55.88	61.71	55.06	84.82	84.85
PgM	**84.69/85.39**	**84.65/85.95**	**66.69**	**66.95**	**73.35**	**70.82**	**90.36**	**91.04**

Table 2: Comparison between modal learning baselines and PgM.

#Model	#Multimodal Sentiment Analysis		#Multimodal Emotion Recognition		#Cross-modal Retrieval		#Image-text Classification	
	MOSI		MELD		Wikipedia		UMPC Food 101	
	ACC-2	**F1**	**ACC**	**Weighted F1**	**MAP**	**Precison@10**	**ACC**	**Weighted F1**
PgM	**84.69/85.39**	**84.65/85.95**	**66.69**	**66.95**	**73.35**	**70.82**	**90.36**	**91.04**
-w/o Modal Partitioner	48.58/50.69	44.31/47.46	51.44	52.72	60.90	59.18	59.69	53.14
-w/o Uni-Modal Learner ($\mathcal{L}^{UFC}$)	65.86/67.18	65.97/67.06	51.75	54.06	63.23	63.64	70.12	70.65
-w/o Paired-Modal Learner ($\mathcal{L}^{PFC}$)	68.14/70.05	66.31/68.69	54.46	56.16	64.68	65.25	73.65	72.96
-w/o Uni-Paired Decoder ($\mathcal{L}^{UPR}$)	69.90/70.69	66.37/67.44	56.81	57.32	64.21	64.64	74.89	74.77

Table 3: Ablation study on various datasets and tasks.

F1 score of 66.95, significantly outperforming baselines, particularly single-modal approaches. Also, in both cross-modal retrieval and image-text classification, PgM consistently achieves superior performance across multiple metrics and tasks, demonstrating its robustness and adaptability in multiple multimodal tasks. These findings highlight the advantages of PgM in two key aspects: (1) effective multimodal learning and (2) enhanced utilization of multimodal information, providing valuable insights for advancing multimodal learning.

	#Model	ACC-2(ACC) %	F1(Weighted F1)%
MSA	Self-MM	84.00/85.98	84.42/85.95
	Self-MM†	85.54/86.75↑	85.18/86.36↑
	MMIM	84.14/86.06	84.00/85.98
	MMIM†	85.36/86.63↑	85.27/86.15↑
MERC	UniMSE	65.09	65.51
	UniMSE†	66.48↑	67.06↑
	UniMEEC	67.36	68.09
	UniMEEC†	67.92↑	68.76↑

Table 4: Adapt PgM to the prior models for MSA on MOSI dataset and MERC on MELD dataset.

4.5 Ablation Study

Table 3 presents an ablation study across multiple datasets and tasks, assessing the impact of the modal partitioner, uni-modal learner, paired-modal learner, and uni-paired modal decoder on model performance. The ablation process entails removing these modules along with their corresponding loss terms from the training objective.

Initially, we remove the modal partitioner from PgM. The uni-modal encoder, paired-modal encoder, and uni-paired modal decoder are built upon the modal partitioner. Consequently, removing the modal partitioner also eliminates the uni-modal encoder, paired-modal encoder, and uni-paired modal decoder, resulting in a significant performance drop, with accuracy (ACC) decreasing to 48.58/50.69 and weighted F1 dropping to 44.31/47.46. Next, we sequentially removed the uni-modal, paired-modal and uni-paired decoder to assess its role. Removing modal learners including uni-modal and paired-modal results in decreased metrics across all tasks. Specifically, in multimodal sentiment analysis and multimodal emotion recognition, removing the modal partitioner and modal learner resulted in reductions of at least 20% and 10% in all metrics, respectively. In cross-modal retrieval, the absence of these components led to a decrease in MAP and Precision@10. In the image-text classification task (UMPC Food 101), removing these components caused drops in ACC and Weighted F1 scores. These declines underscore the effectiveness of combining both modal partitioner and modal learner in multimodal learning. In conclusion, each module in PgM play a crucial role in significantly improving model performance across various tasks and their corresponding metrics.

4.6 Adapting PgM for Multimodal Tasks

We further evaluate the adaptivity of PgM on existing multimodal models, i.e., we enhance the performance of existing models with PgM. The results (enhanced performance is indicated by an underline) are given in Table 4. We integrate the proposed PgM framework into well-established models (e.g., Self-MM, MMIM, UniMSE, and UniMEEC) in the MSA and MERC fields, with their enhanced

versions denoted by the superscript †. Specifically, we integrate our proposed PgM into their framework by replacing their multimodal learning modules. For MSA task, Self-MM† demonstrates a substantial improvement, achieving 85.54/86.75 ACC-2 and 85.18/86.36 F1 score, outperforming original Self-MM's results, highlighting PgM's ability to enhance multimodal learning. Similarly, MMIM† also demonstrates improvement, achieving ACC-2 scores of 85.36/86.63 and F1 scores of 85.27/86.15, with approximately 0.5–1% gains across all metrics. However, it benefits less from the proposed multimodal learning framework compared to other models. For the MERC task, UniMSE† shows the most significant improvement, reaching 66.48 ACC and 67.06 Weighted F1, with an approximate 1–2% boost across all metrics. Similarly, UniMEEC† achieves the highest gains, with 67.92 ACC and 68.76 Weighted F1. These results confirm PgM's effectiveness in enhancing already strong models and further highlight its adaptability.

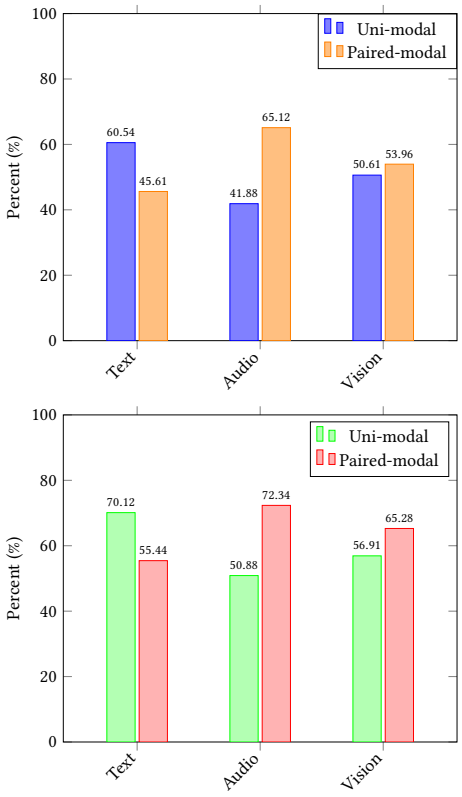

Figure 4: Distribution of uni-modal and paired-modal features across text, audio and vision modalities for MSA (top) and MERC (bottom).

4.7 Visualization

Furthermore, we visualize the distribution of uni-modal and paired-modal features for each individual modality after applying the modal partitioner, as shown in Figure 4.

In the text modality, 60.54% of features are uni-modal, and 45.61% are paired-modal, indicating that uni-modal features independently provide more information for the MSA task. In the audio modality, 41.88% are uni-modal and 65.12% are paired-modal, suggesting that its paired-modal features play a larger role in the MSA task. In the vision modality, uni-modal and paired-modal features are nearly evenly distributed (50.61% vs 53.96%), indicating that both uni-modal and paired-modal features contribute to MSA task. Similarly, we analyze the distribution of uni-modal and paired-modal features across modalities in the MERC task, which differs from the MSA task. For example, in MERC, uni-modal features in the text modality remain the dominant source of information, while paired-modal features in the audio and vision modalities contribute more compared to their respective uni-modal features. We have two observations on the visualizations: 1)uni-modal and paired-modal features exhibit a certain degree of overlap across different modalities and tasks, and 2)the modal partitioner can adjust the distribution of uni-modal and paired-modal features across different multimodal tasks.

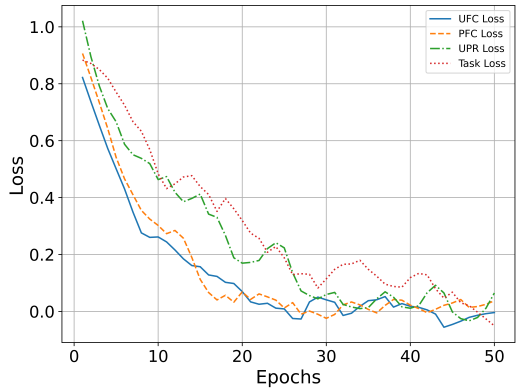

Figure 5: Loss variation curves during the second stage, i.e., the joint training process of the downstream task and PgM for MSA task.

4.8 PgM Training Loss on a Downstream Task

Figure 5 illustrates the loss variation curves for the joint training process of the downstream task with PgM. In the early training stages (first 10 epochs), the losses decrease rapidly, suggesting that the model quickly captures essential features. UFC Loss and PFC Loss exhibit a steep decline and stabilize after approximately 20 epochs, implying that these components learn efficiently. UPR Loss and Task Loss decrease at a slower rate and exhibit fluctuations even after 20 epochs, suggesting that these losses are influenced by more complex modality interactions. By around epoch 30, all loss terms stabilize and approach zero, indicating that the model has largely converged. However, between epochs 30 and 50, Task Loss and UPR Loss still exhibit fluctuations. These loss curves reveal that UFC, PFC, UPR, and Task Loss decrease at different rates, highlighting PgM's varying learning dynamics across uni-modal feature, paired-modal feature, and downstream tasks. Moreover, these results suggest that PgM alleviates the modality laziness during training that is commonly observed in previous multimodal learning work [4].

4.9 Model Size

To further improve the understanding of PgM, we quantify the training parameter sizes for each module. The sizes of trainable parameters for each module are reported in Table 5. During the training phase, the parameters of the modal encoders remain fixed, while only those in the modal partitioner, uni-modal learner, paired-modal learner, and uni-paired decoder are updated based on the training loss. The modal partitioner requires only a small number of trainable parameters, highlighting its lightweight nature in feature partitioning. In contrast, the uni-modal learner, paired-modal learner, and uni-paired decoder adopt a Transformer-based architecture, each containing approximately 7.1M trainable parameters. While PgM incorporates more parameters than prior multimodal learning frameworks—attributable to the inclusion of the partitioner training module—the overall parameter increase remains within a tolerable range.

Modules	Trainable Parameter
Modal encoder	0M
Modal partitioner	1.1M
Uni-modal learner	7.08M
Paired-modal learner	7.08M
Uni-Paired decoder	7.1M
PgM	22.36M

Table 5: The number of trainable parameters.

5 Conclusion

This paper presents PgM, a partitioner-guided modal learning framework consisting of the modal partitioner, modal learner, and uni-paired modal decoder. The modal partitioner divides the modal representation into uni-modal and paired-modal features, the modal learner enhances their learning through dedicated uni-modal and paired-modal components, and the uni-paired decoder reconstructs the modal representation. Extensive experiments on four multimodal tasks—spanning vision, language, and audio modalities—demonstrate PgM's versatility and effectiveness. Additionally, we visualize the contributions of uni-modal and paired-modal features to multimodal tasks, offering valuable insights into their respective roles. We believe this work presents a new experimental setting that can provide a new and different perspective to multimodal learning communities.

Acknowledgement

We sincerely thank all coauthors for their contributions. This work was supported by research grants from VILLUM FONDEN (VIL50296).

References

[1] Relja Arandjelovic and Andrew Zisserman. 2017. Look, Listen and Learn. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 609–617. doi:10.1109/ICCV.2017.73

[2] Shunjie Chen, Xiaochuan Shi, Jingye Li, Shengqiong Wu, Hao Fei, Fei Li, and Donghong Ji. 2022. Joint Alignment of Multi-Task Feature and Label Spaces for Emotion Cause Pair Extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*. 6955–6965.

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

[4] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On Uni-Modal Feature Learning in Supervised Multi-Modal Learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 8632–8656. https://proceedings.mlr.press/v202/du23e.html

[5] Yuan Gong, Yu-An Chung, and James R. Glass. 2021. AST: Audio Spectrogram Transformer. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek (Eds.). ISCA, 571–575. doi:10.21437/INTERSPEECH.2021-698

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 6325–6334. doi:10.1109/CVPR.2017.670

[7] Nitin Grover, Aviral Chharia, Rahul Upadhyay, and Luca Longo. 2023. Schizo-Net: A novel Schizophrenia Diagnosis Framework Using Late Fusion Multimodal Deep Learning on Electroencephalogram-Based Brain Connectivity Indices. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2023), 464–473. doi:10.1109/TNSRE.2023.3237375

[8] Wei Han, Hui Chen, and Soujanya Poria. [n. d.]. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). 9180–9192.

[9] Guimin Hu, Daniel Hershcovich, and Hasti Seifi. 2025. HapticLLaMA: A Multimodal Sensory Language Model for Haptic Captioning. *arXiv preprint arXiv:2508.06475* (2025).

[10] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256* (2022).

[11] Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Lijie Hu, Hasti Seifi, and Jiayuan Xie. 2024. Unimeec: Towards unified multimodal emotion recognition and emotion cause. *arXiv preprint arXiv:2404.00403* (2024).

[12] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What Makes Multi-Modal Learning Better than Single (Provably). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 10944–10956. https://proceedings.neurips.cc/paper/2021/hash/5aa3405a3f865c10f420a4a7b55cbff3-Abstract.html

[13] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning? (Provably). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 9226–9259. https://proceedings.mlr.press/v162/huang22e.html

[14] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2023. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7661–7671.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[18] Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, et al. 2024. FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 19077–19095.

[19] Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. 2024. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion

recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*. 41–48.

[20] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2022. MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. *CoRR* abs/2207.00056 (2022). arXiv:2207.00056 doi:10.48550/ARXIV.2207.00056

[21] Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-SENA: An integrated platform for multimodal sentiment analysis. *arXiv preprint arXiv:2203.12441* (2022).

[22] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. 2019. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6966–6975.

[23] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 527–536. doi:10.18653/v1/p19-1050

[24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[25] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. 251–260.

[26] Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2018. Ordered neurons: Integrating tree structures into recurrent neural networks. *arXiv preprint arXiv:1810.09536* (2018).

[27] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15638–15650.

[28] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 6558–6569. doi:10.18653/v1/p19-1656

[29] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[30] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What Makes Training Multi-Modal Classification Networks Hard?. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 12692–12702. doi:10.1109/CVPR42600.2020.01271

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

[32] Jiayuan Xie, Mengqiu Cheng, Xinting Zhang, Yi Cai, Guimin Hu, Mengying Xie, and Qing Li. 2025. Explicitly Guided Difficulty-Controllable Visual Question Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 25552–25560.

[33] Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A Partition Filter Network for Joint Entity and Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 185–197.

[34] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 10790–10797.

[35] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intell. Syst.* 31, 6 (2016), 82–88. doi:10.1109/MIS.2016.94

[36] Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor Versatile Multi-Modal Learning for Multi-Label Emotion Recognition. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 9100–9108. doi:10.1609/AAAI.V36I8.20895

[37] Zhihong Zhu, Xuxin Cheng, Guimin Hu, Yaowei Li, Zhiqi Huang, and Yuexian Zou. 2024. Towards Multi-modal Sarcasm Detection via Disentangled Multi-grained Multi-modal Distilling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 16581–16591.

[38] Zhihong Zhu, Xianwei Zhuang, Yunyan Zhang, Derong Xu, Guimin Hu, Xian Wu, and Yefeng Zheng. 2024. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In *Proc. of IJCAI*.