Actual Cause Guided Adaptive Gradient Scaling for Balanced Multimodal Sentiment Analysis

JILI CHEN and QIONGHAO HUANG*, Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, China
CHANGQIN HUANG*, College of Education, Zhejiang University, China
XIAODI HUANG, School of Computing, Mathematics and Engineering, Charles Sturt University, Australia

Multimodal sentiment analysis leverages information from multiple sensors to achieve a comprehensive interpretation of emotions. However, different modalities do not always boost each other as expected. They compete with each other, leading to some modalities being under-optimized during the training process. To address this issue, we propose **A**daptive **G**radient **S**caling with **S**parse **M**ixture-**o**f-**E**xperts (AGS-SMoE). We first discuss the issue of modal preemption in unified multimodal learning from the perspective of causal preemption. Driven by actual cause, we use the gradient norms from different encoders at two fusion stages as evidence, estimating the current modal preemption state using a parameter-free method. Then, based on the dynamic preemption factor, we design a gradient scaling method to balance optimization for different encoders. Furthermore, we use Mixture-of-Experts to sparsify and perceive multimodal tokens in different preemption states. As a result, our experiments on four multimodal sentiment analysis datasets have achieved state-of-the-art results. Moreover, our method improves modal representation learning at different stages. Extensive experiments confirm that our method can alleviate the modal preemption problem in a plug-and-play manner. Our code is available at https://github.com/TheShy-Dream/AGS-SMoE.

CCS Concepts: • **Information systems** → **Sentiment analysis**.

Additional Key Words and Phrases: Multimodal Sentiment Analysis, Causal Preemption, Multimodal Competition.

1 INTRODUCTION

Multimodal sentiment analysis (MSA) is a hotspot field that aims to decode emotional expressions by synergistically interpreting various modalities such as facial expressions, speech, and text [3]. This approach surpasses unimodal methods, offering a more nuanced understanding of emotions [10]. It is applied in diverse scenarios including healthcare for mental state assessment, automated tutoring systems for emotional support, and enhancing customer service through call analytics [71]. Previous research has confirmed that designing more effective fusion networks [47, 56, 59] and incorporating multi-task learning methods [13, 25, 36, 63] that reduce modality gaps can significantly promote multimodal integration. Most MSA studies [15, 17–19, 51] are based on the assumption that in a unified multimodal training, all modalities can be adequately optimized.

*Qionghao Huang (qhhuang@m.scnu.edu.cn) and Changqin Huang (cqhuang@zju.edu.cn) are corresponding authors.

This work was supported by the National Science and Technology Major Project (No. 2022ZD0117104).
Authors' Contact Information: Jili Chen, irelia@zjnu.edu.cn; Qionghao Huang, qhhuang@m.scnu.edu.cn, Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua, China; Changqin Huang, cqhuang@zju.edu.cn, College of Education, Zhejiang University, Hangzhou, China; Xiaodi Huang, xhuang@csu.edu.au, School of Computing, Mathematics and Engineering, Charles Sturt University, Albury, Australia.

ACM 1551-6865/2025/5-ART
https://doi.org/10.1145/3736415

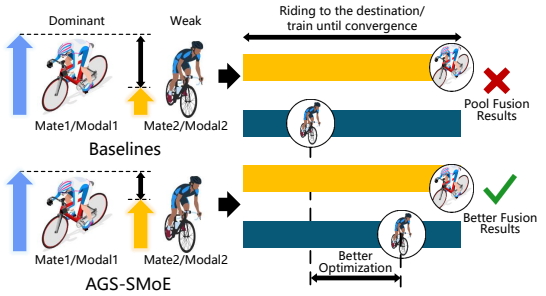
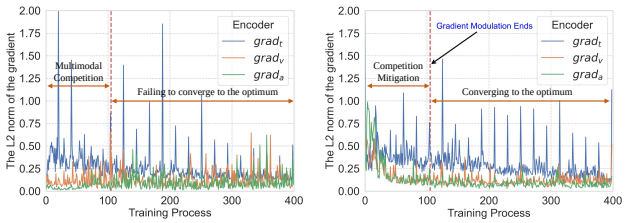


Fig. 1. A vivid example of multimodal cooperation and competition. In a team cycling event, drafting and cooperation among team members help enhance the overall capability of the team. The same principle applies to the collaboration among different modalities in a multimodal context. By employing a multimodal collaborative strategy, AGS-SMoE can achieve better fusion results.



(a) Training process without gradient scaling.

(b) Training process with AGS-SMoE.

Fig. 2. The average L2 norm of the gradients for different encoders in a joint multimodal training process on the CMU-MOSI dataset.

However, most deep learning models are based on likelihood maximization, and optimizing better-performing modalities brings the fastest convergence of discriminative loss. But it may also dominate the training process, causing competition between modalities and leading to the under-optimization of other modalities (also referred to as modality laziness [5, 68]). As shown in Fig. 2(a), the average L2 gradient norm of the text encoder converges significantly throughout the training process, while the convergence of the audio and visual encoders is not obvious. That is because the text contains most sentimental information and is quickly optimized, while the optimization process of other modalities is suppressed by the optimization of the text encoder. The fusion result is generated through the interaction of two modalities, and it is unreasonable to optimize just one of them. This will compromise the model's interpretability and generalization ability. In Fig. 1, we compare the different modalities in a multimodal task to different team members: Typically, the stronger modalities (like stronger cyclists) lack connection and mutual assistance with the weaker modalities (slower cyclists). Even if they reach the finish line first (achieve training convergence), they have to wait for the weaker modalities to catch up because the team's final time is determined by when their last member crosses the finish line (the comprehensive performance of all modalities). Through appropriate regulation, the stronger members can slow down to help the weaker ones accelerate more effectively. The weaker members behind them can increase their speed thanks to this assistance (drafting or other collaborations), thereby improving the team's overall performance. Collaboration and mutual assistance among modalities can alleviate the suboptimal issues arising from multimodal competition.

Recently, studies have focused on the issue of modal competition, some [22, 68] utilize teacher models to enhance the learning of a single modality through distillation, others [26, 46, 57] use gradient modulation methods to balance optimization of different modalities. However, these methods still suffer from the following issues: 1) The traditional methods only evaluate multimodal competition state based on the inconsistency of accuracy among different modality encoders and forcibly require the optimization states of all modalities to be aligned, ignoring the inherent causality of the modalities on the fusion results which often leads to a degradation in multimodal fusion. 2) When more than two modalities are fused, the competition among modalities becomes more complex. Existing methods struggle to generalize to complex situations involving multiple dominant or suppressed modalities.

In response to the aforementioned issues, it naturally raises the question: ***Q1: Is it possible to use a causal perspective to assess multimodal preemption status?*** Gradient-based modal competition mitigation can be

regarded as regularization. The assessment of modality preemption status is a prerequisite for gradient scaling and directly affects the model performance. However, due to the black-box nature of neural networks, effectively identifying the dominant modality from a causal perspective is challenging. Even if the dominant modality is successfully identified, we still face another question: ***Q2: How to design appropriate gradient scaling strategies based on the preemption intensity when involving multiple dominant or suppressed modalities?*** For different situations across various modalities, gradients should be dynamically adjusted based on the modality preemption status to balance the optimization of different modalities and maximize the advantages of each modality.

To address the aforementioned questions, we embrace actual cause theory and propose **A**daptive **G**radient **S**caling with **S**parse **M**ixture-**o**f-**E**xperts (AGS-SMoE), which addresses the issue of modal competition from the perspective of causal preemption. To answer ***Q1***, we use the L2 norm of gradients from different encoders in each iteration to dynamically estimate the actual cause of different modalities on the fusion result. Based on the actual causal effect, we further assess the preemptive modality and preemption factor for each batch. To answer ***Q2***, we set unique gradient update strategies for each encoder based on the preemption factor to dynamically control the optimization speeds of different modalities and alleviate the problem of modal competition. We have adopted the aforementioned gradient scaling strategy for different fusion stages. Regarding token-level preemption, we utilize a Mixture of Experts (MoE) [48] as the fusion network to dynamically allocate multimodal tokens according to the varying degrees of preemption. Recent advancements in MoE have demonstrated their effectiveness in enhancing model performance [38, 52]. While prior approaches [31, 32, 49] primarily focus on architectural enhancements to MoE, our method directly leverages MoEs to manipulate gradient flows for balanced optimization across different modalities. As shown in Fig. 2(b), our AGS-SMoE can alleviate multimodal competition and ensure that all modalities are well-optimized. It is worth mentioning that AGS-SMoE is a parameter-free and model-agnostic method. The contributions of this paper are as follows:

- We propose AGS-SMoE, a parameter-free and model-agnostic method, which effectively alleviates multimodal competition in multimodal joint training through multimodal causal preemption state estimation and adaptive gradient scaling strategies.
- We are the first work to systematically analyze the problem of multimodal competition using causal preemption theory and provide theoretical support for preemption state estimation. The theory proves that our method is applicable to more modalities and more stages of the training process.
- Our experiments on the four open datasets for multimodal sentiment analysis CMU-MOSI [66], CMU-MOSEI [67], UR-FUNNY [14], and CH-SIMS [62] show that AGS-SMoE has achieved new state-of-the-art (SOTA) results. The extensive experiments confirm that AGS-SMoE not only achieves better fusion results but also promotes representation learning at different fusion stages.

2 RELATED WORK

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis goes beyond traditional text analysis by considering visual cues and potentially auditory inputs to provide a more accurate sentiment assessment. To effectively aggregate consistent information from different modalities, modal fusion is a core issue in multimodal sentiment analysis. Zedeh *et al.* [64] proposes a tensor-based network for multi-level multimodal fusion. For better sequential learning, Zedeh *et al.* [65] models view-specific and cross-view interactions through a system of LSTMs. Many researchers have incorporated self-supervised learning [15, 63] and mutual information [13] into tensor fusion networks to reduce the burden of multimodal learning. The breakthroughs of the Transformer [55] in the fields of natural language processing and computer vision have also enhanced the development of cross-modal interaction. Tsai *et al.* [53] leverage the ability of the Transformer to learn sequential weight mappings, extending it to a cross-modal Transformer for

multimodal sentiment analysis. Some researchers [35, 56] have also noted the unique role of text in multimodal sentiment analysis and have designed text-centric attention networks. Kim *et al.* [25] integrate fusion within the pre-trained Large-BERT, while Hu *et al.* [17] place the fusion process within the pre-trained T5 model, demonstrating that language model-based large-scale fusion can improve accuracy. On the other hand, more and more auxiliary tasks are designed to reduce the modal gap and enhance accuracy [9, 29, 30, 34].

All the aforementioned MSA models implement joint training within a uniform network, failing to account for the competitive relationships between modalities, resulting in unimodal representation degradation. Therefore, we analyze and integrate the issue of suboptimal modality optimization caused by joint training, and design a gradient scaling method to mitigate this issue.

2.2 Multimodal Competition

When different modalities are integrated to obtain consistent information, there is also interference between modalities, which makes the joint training process suboptimal. To explain the causes of modality competition during joint training, Huang *et al.* [21] provide both empirical evidence and proof at the experimental level. Wang *et al.* [57] find that multi-modal networks often underperform compared to the best uni-modal networks, they compute an optimal blend of modalities and use gradient-blending to tackle these problems. Du *et al.* [5] combine fusion objectives with uni-modal distillation and introduce Uni-Modal Teacher (UMT) to learn sufficient features from each modality. To make better use of visual information and enhance the balance between the visual and textual modalities, Winterbottom *et al.* [58] evaluate the contributions of visual and textual information and propose bilinear pooling fusion. Peng *et al.* [46] propose OGM-GE which controls the optimization of each modality by monitoring the discrepancy in their contributions to the learning objective and introduces extra Gaussian noise that changes dynamically to counteract any potential drop in generalization ability caused by the gradient modulation. Li *et al.* [26] introduce an Adaptive Gradient Modulation (AGM) method to boost the performance of multi-modal models by dynamically adjusting the gradient signals during back-propagation. Zhang *et al.* [69] developed gradient coordination to alleviate multi-modal gradient conflicts.

These methods control the balance of modalities in competitive relationships, mostly based on the enhancement of encoders or gradient modulations. They only consider the competition among different modalities through the assessment of accuracy and forcibly require different modalities to align the optimization states together, which can also harm model performance. When the number of modalities increases, it becomes very difficult to enforce consistent optimization across all modalities, and the model often struggles to converge. Most of them require additional training parameters to enhance the representational capacity of specific modalities. Unlike previous methods, AGS-SMoE aims to align the optimization with the actual cause estimated by modality-specific gradients. Due to the objective stability of the inherent causality of the modalities, using it to scale gradients is more stable than relying on unimodal accuracy. The gradient estimation approach allows AGS-SMoE to operate without additional unimodal classifiers. It can boost balanced optimization in a parameter-free and model-agnostic manner from the perspective of causal preemption.

3 PROBLEM FORMULATION

In this section, we introduce the task definitions, basic concepts, and causal assumptions which will be used for theoretical analysis in Section 4.

Task Definition. The objective of traditional multimodal sentiment analysis (MSA) is to develop a computational model that can accurately classify or predict the sentimental state conveyed by a combination of modalities, such as text, audio, and visual data [18]:

$$Y_i = f_\theta(I_t \in \mathbb{R}^{l_t \times d_t}, I_a \in \mathbb{R}^{l_a \times d_a}, I_v \in \mathbb{R}^{l_v \times d_v}), \tag{1}$$

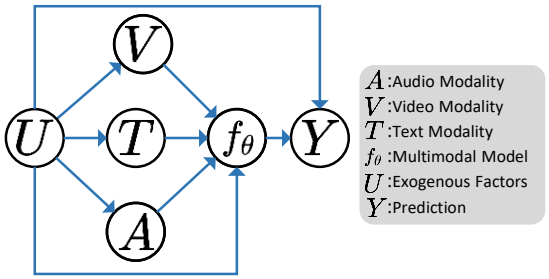


Fig. 3. Causal graph of multimodal sentiment analysis.

where f_θ is the function that maps the input modalities to a sentimental state. I_t represents text data, which could include linguistic features. I_a represents audio data, which might involve acoustic features like pitch, tone, and intensity. I_v represents visual data, encompassing facial expressions, body language, or other visual cues. The lengths of the text data, audio data, and visual data are l_t, l_a, and l_v, respectively. Their dimensions are d_t, d_a, and d_v, respectively.

Our objective is to tackle the issue of multimodal competition in MSA. The goal is to learn a mapping $f_\theta^* \in f_\theta$ ensuring not only a good final fusion representation but also good representations for different modality branches within a MSA model.

Definition 3.1 (Causal Graph). The causal graph is a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ to represent the causal relationships between a set of variables. It consists of nodes that represent the variables $\mathcal{N}$ and directed edges $\mathcal{E}$ that indicate the direction of causality from cause to effect [44]. Fig. 3 presents a causal graph of MSA.

Definition 3.2 (Causal Model) [45]. A causal model can be represented by the following triplet:

$$M = \langle U, V, F \rangle, \tag{2}$$

where U denotes the set of exogenous variables, whose values are determined by external factors. V denotes the set of endogenous variables, whose values are determined by their parent node set PA, and F represents the mapping relationships between variables. Each model M is associated with a causal graph, where each node $\mathcal{N}$ in the causal graph corresponds to an endogenous or exogenous variable, and the edges $\mathcal{E}$ in the causal graph correspond to the mapping relationships between the nodes. In Fig. 3, U represents the exogenous variable set, while all other variables are endogenous. The blue lines indicate the mapping relationship F. It is worth mentioning that the exogenous variable U will affect the mapping relationship F, thereby causing changes in some endogenous variables V. For instance, different sampling orders or varying computational quantization levels can impact the model and affect the final prediction.

Definition 3.3 (Causal Beam) [45]. For model $M = \langle U, V, \{f_i\} \rangle$ and state $U = u$, a causal beam is a new model $M_u = \langle u, V, \{f_i^u\} \rangle$, in which the set $\{f_i^u\}$ is constructed through the following steps:

(1) For each variable $V_i \in V$, its parent nodes in set PA_i are divided into two subsets $PA_i = S \cup \bar{S}$, such that no matter how we set the members of $\bar{S}$, S is any subset of PA_i that is sufficient to imply the actual value of $V_i(u)$:

$$f_i(S(u), \bar{S}, u) = f_i(S(u), \bar{S}', u) \quad \forall \bar{s}'. \tag{3}$$

(2) For each variable $V_i \in V$, identify a set $X \subset \bar{S}$ such that certain realizations $X = x$ can make the function $f_i(s, \bar{S}_x(u), u)$ nontrivial in s:

$$f_i(s', \bar{S}_x(u), u) \neq V_i(u) \quad \exists s'. \tag{4}$$

(3) Replace $f_i(s, \bar{s}, u)$ with the mapping $f_i^u(s)$:

$$f_i^u(s) = f_i(s, \bar{S}_x(u), u). \tag{5}$$

The causal beam provides us with a fundamental environment for analyzing causal preemption, on the basis of which we describe the Natural Causal Beam.

Definition 3.4 (Natural Causal Beam) [45]. M_u is a Natural Causal Beam when the condition 2 of the definition 3.3 is satisfied with $X = \emptyset$, meaning all x variables maintain their true values instead of being freezed through $do(\cdot)$. At this point, step 3 can be simplified to $f_i^u(s) = f_i(s, \bar{S}(u), u)$. Once a Natural Causal Beam exists, we can identify the actual cause of the event within the current beam.

Definition 3.5 (Actual Cause) [45]. Event $X = x$ is the actual cause of $Y = y$ if and only if the following conditions are met within the natural beam M_u:

$$Y_x = y \ in \ M_u \ and \ Y_{x'} \neq y \ in \ M_u \quad \exists x' \neq x. \tag{6}$$

The state u is uncertain and we use $P(u)$ to represent its probability. $P(u)$ also reflects the probability that x is the actual cause of y, represented as cause(x, y). The state u can be estimated based on our observational evidence e, denoted as $P(u|e)$. Additionally, we use U_{xy} to denote the subset of states where x is the actual cause of y, and U_e to denote the subset of states that are consistent with the observed evidence e, Thus far, we have obtained an estimation method for the probability that x is the actual cause of y given the evidence [45]:

$$P(\text{cause}(x, y)|e) = \frac{P(U_{xy} \cap U_e)}{P(U_e)} = P(u = 1|e). \tag{7}$$

4 METHODOLOGY

In this section, we first analyze the modality competition issue from the perspective of Multimodal Causal Preemption (Section 4.1.1). Using Causal Beams as a tool to estimate the preemption status (Section 4.1.2), we then present the comprehensive architecture of our proposed AGS-SMoE (Section 4.2). Finally, we discuss the mechanism of AGS-SMoE (Section 4.3).

4.1 Theoretical analysis of multimodal competition

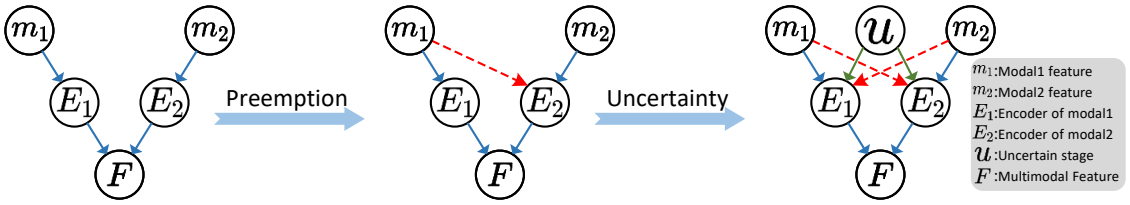
In this section, we refer to Structural Causal Model (SCM) [45] to represent the competition of modal representations as a causal preemption problem and introduce the relevant causal relationships [6]. Our subsequent methods will be based on these causal relationships and assumptions.

4.1.1 Multimodal Causal Preemption. Based on the process of traditional multimodal fusion, we construct the causal graph as shown in Fig. 4(a). The structural equations are not explicitly depicted in the figure, yet they are assumed to ascertain the value of each child variable based on the values of its parental variables within the graph. m_1 and m_2 represent the two modalities to be fused, E_1 and E_2 represent the corresponding modality encoders, and F represents the multimodal feature after fusion. We can utilize the following boolean model to describe the multimodal fusion in the ideal state [6]:

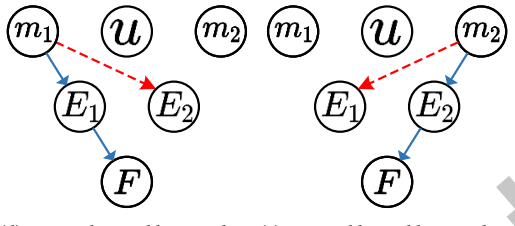
$$e_1 = m_1, \quad e_2 = m_2, \quad f = e_1 \vee e_2 = m_1 \vee m_2, \tag{8}$$

the lowercase letters represent the output of the corresponding modules, for example, $e_1 = m_1$ indicates that the modality m_1 is mapped by a structural equation (estimated through the encoder E_1) to obtain e_1. It is worth mentioning that the expressions on either side of the equals sign here describe a causal relationship, thus they are not interchangeable.

However, most multimodal sentiment analysis models are based on the premise that the encoding processes of each modality are independent during joint training. Many studies [21, 26, 46] have indicated that within a



(a) Ideal multimodal fusion. (b) Causal preemption in multimodal fusion. (c) Causal preemption in multimodal fusion with uncertainty.



(d) Natural causal beam when state u=1. (e) Natural bausal beam when state u=0.

Fig. 4. The causal preemption in multimodal fusion depicted using causal graphs. Blue solid lines represent the ideal causal transmission edges, red dashed lines indicate causal preemption edges caused by modal competition, and green solid lines illustrate complex causal preemption under the influence of uncertain factors. Fig. 4(a), Fig. 4(b), and Fig. 4(c) describe process for causal preemption in multimodal learning. The Fig. 4(d) and Fig. 4(e) illustrate the schematic diagram of natural causal beams for two different states.

unified training framework, modalities with better performance will dominate the entire training process, leading to other modalities being sub-optimized.

In causal graphs, we employ the concept of causal preemption to depict this issue of modality competition [39], which represents a more generalized scenario [45]. Taking the example of m_1 preempting m_2, we illustrate this state with Fig. 4(b). We can use the following formula to describe the state:

$$e_1 = m_1, \quad e_2 = m_1' \wedge m_2, \quad f = e_1 \vee e_2, \tag{9}$$

when we import e_1 and e_2 into the expression of f, we can obtain a simple disjunction as follows:

$$f = m_1 \vee (m_1' \wedge m_2) \equiv m_1 \vee m_2, \tag{10}$$

the significance of structural information is underscored in this formula. This also characterizes the gap between multimodal competition and the ideal of multimodal learning. The two sides of the equation are logically equivalent, yet their structures are not symmetrical [2]. $m_1 \vee m_2$ maintains symmetry between m_1 and m_2, allowing for interchangeability, while $m_1 \vee (m_1' \wedge m_2)$ indicates that m_2 has an effect on f when only m_1 is not true. When m_1 is true, m_2 does not affect on the outcome f. This aligns with the objective of multimodal training: when the primary modality m_1 does not possess specific information, m_2 provides auxiliary information for predictions. However, the actual situation is that the model extracts more information in m_1, and thus prioritizes the optimization of e_1, only optimizing e_2 when m_1 is devoid of information. This leads to e_2 being sub-optimized, diminishing its capacity to extract information from m_2. It creates a vicious cycle where, even if there are sentimental cues within m_2, they can not extract useful information due to the consistently sub-optimized state of the encoder. Such failure further worsens the sub-optimized state of the encoder and impairs the entire network.

During the training process, the roles of primary and secondary modalities are determined by the batch distribution and the current optimization state of the encoder. Therefore, in the dynamic training process, each modality has a probability of becoming a primary modality. If a modality becomes the primary one, it is more likely to preempt other modalities in the current iteration. We model this probabilistic process as shown in Fig. 4(c). To model this uncertainty, we introduce a binary variable u to represent whether modality m_1 preempts m_2 ($u = 1$) or m_2 preempts m_1 ($u = 0$). To simplify the causal model, we do not introduce background variables U_{m_1} and U_{m_2} for m_1 and m_2, which will not affect the conclusions we draw [7, 45]. We can derive the following formula:

$$e_1 = m_1 \wedge (u' \vee m_2'), \quad e_2 = m_2 \wedge (u \vee m_1'), \quad f = e_1 \vee e_2, \tag{11}$$

this is a more generalized case of Eq. 9, when $u = 1$, it simplifies to Eq. 9. The state u directly influences the causal effects between different modalities and the results of modal fusion.

In the multimodal learning framework, u is not a static binary variable but a dynamic value. Because the state of modality preemption is closely related to the data distribution of the dataset and the characteristics of the modalities. This probability-based single-event causation requires the application of the Causal Beam criterion for modeling [2, 43].

4.1.2 Preemption State Estimation in Multimodal Learning. In multimodal joint training, when we set $u = 1$ in Eq. 11, we can obtain Eq. 9. According to Definition 3.3, we can identify the following sustaining parent sets: m_1 (for E_1), m_1 (for E_2), E_1 (for Y), thus, the projection in causal beam model $M_{u=1}$ is as follows:

$$e_1 = m_1, \quad e_2 = m_1', \quad f = e_1, \tag{12}$$

furthermore, we find that at this point, the variables in the complement of the sustaining parent set for each variable only need to maintain their original values to achieve $F = f'$ by setting $m_1 = m_1'$. Therefore, $M_{u=1}$ is a natural causal beam according to Definition 3.4, and its structure is shown in Fig. 4(d). According to Eq. 6, we obtain

$$\begin{aligned} F_{m_1} = f \quad and \quad F_{m_1'} = f' \quad in \quad M_{u=1}, \\ F_{m_2} = f \quad and \quad F_{m_2'} = f \quad in \quad M_{u=1}, \end{aligned} \tag{13}$$

it is evident that when $u = 1$, m_1 establishes a counterfactual dependence with the fusion result, and therefore m_1 is the actual cause of F, whereas m_2 is not the actual cause of F.

On the contrary, when $u = 0$, the the projection in causal beam model $M_{u=0}$ becomes:

$$e_1 = m_2', \quad e_2 = m_2, \quad f = e_2, \tag{14}$$

which is also a natural beam as shown in Fig. 4(e), and the test formula is:

$$\begin{aligned} F_{m_1} = f \quad and \quad F_{m_1'} = f \quad in \quad M_{u=0}, \\ F_{m_2} = f \quad and \quad F_{m_2'} = f' \quad in \quad M_{u=0}, \end{aligned} \tag{15}$$

it is clear that when $u = 0$, m_2 is the actual cause of the fusion result, while m_1 is not. According to Eq. 6, we obtain:

$$P(cause(m_1, F|e)) = P(u = 1|e), \quad P(cause(m_2, F|e)) = P(u = 0|e), \tag{16}$$

when extending to more modalities, we use the $\tilde{u}_{m_1}$ to represent the preemptive capability of the modality m_i, which also reflects the extent to which a specific modality is the actual cause of the fusion result [45]:

$$P(cause(m_i, F|e)) = P(u = \tilde{u}_{m_i}|e). \tag{17}$$

4.2 AGS-SMoE

Based on the aforementioned theoretical analysis, we propose AGS-SMoE. Its architecture primarily consists of unimodal feature extraction and cross-modal MoEs. The core method involves using deep neural networks to estimate causal preemption states and to perform adaptive gradient scaling to mitigate modality competition. The overall pipeline of AGS-SMoE is shown in Fig. 5.

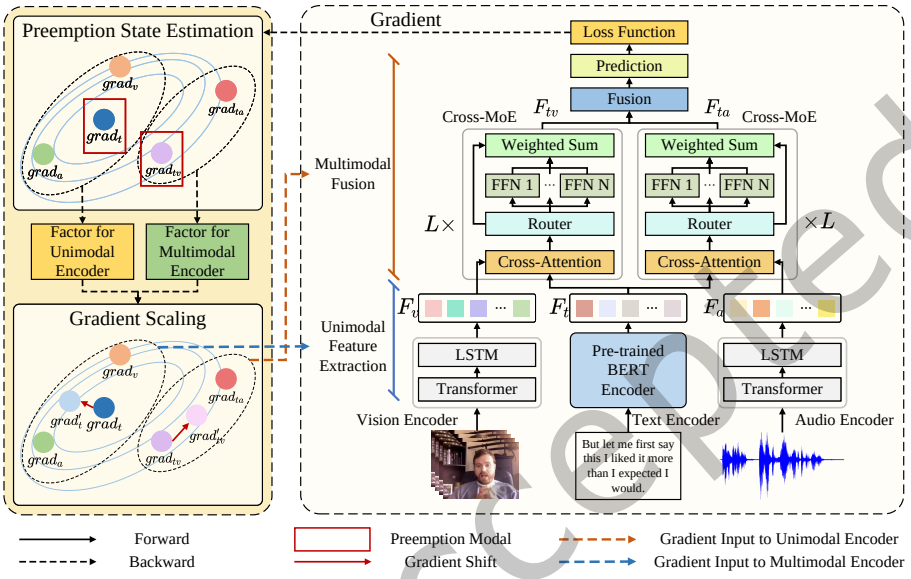


Fig. 5. **Pipeline of AGS-SMoE.** The pipeline mainly consists of four parts: unimodal feature extraction, multimodal fusion, preemption state estimation and gradient scaling. In the unimodal feature extraction and multimodal fusion, we estimate the multimodal preemption by the gradient norm, identify the dominant modality, and estimate the degree of preemption in a parameter-free manner, adjusting the gradient according to the varying degrees of dominance. We use two Mixture-of-Experts (MoE) models as our fusion network, where different experts are encouraged to learn the multimodal token features under different dominant conditions and obtain the fusion result through weighted summation.

4.2.1 Model Architecture. **Unimodal Feature Extraction.** Following [13, 63], for the text modality, we use BERT_{base} as the text encoder to map the text input into a sequence of vectors. We use Facet [40] for visual feature extraction and COVAREP [4] for audio spectral feature extraction. Then the audio and visual features are processed by Transformers and unidirectional LSTM [16] networks, respectively:

$$F_t = BERT(I_t), \quad F_m = LSTM(Transformer(I_m)), m \in \{a, v\}. \tag{18}$$

Multimodal Fusion with Cross-MoE. The Mixture-of-Experts (MoE) [48] has achieved new breakthroughs in both large language models [23, 60, 61] and large multimodal models [27, 28, 31, 42, 50]. We utilize MoE for two main reasons: 1. MoE enables each expert to specialize in handling different types of multimodal tokens, enhancing the model's efficiency by alleviating the computational load on individual experts and allowing for a more concentrated data analysis [28, 31, 42]. 2. When we apply gradient scaling strategies, MoE can assign different experts to tokens based on their varying preemption states through conditional gating, thereby encouraging diverse experts to learn the integration methods under different preemption conditions. Such integration methods

can better align with our gradient scaling strategy. We utilize a L-layer Cross-modal Noisy Gating Mixture-of-Experts (Cross-MoE) as our fusion backbone network. The Cross-MoE consists of a cross-modal attention module [53] followed by a MoE module [48]:

$$\begin{aligned} h_{tm}^{i} &= Attention(Q = h_{t}^{i}, K = h_{m}^{i}, V = h_{m}^{i}), m \in \{a, v\}, \\ H_{n}^{i} &= (h_{tm}^{i} \cdot W_{g}^{i})_{n} + \epsilon \cdot Softplus((h_{tm}^{i} \cdot W_{noise}^{i})_{n}), m \in \{a, v\}, \\ h_{tm}^{i+1} &= \sum_{n=1}^{N} Softmax(Top(H^{i}, k))_{n} FFN_{n}^{i}(h_{tm}^{i}), m \in \{a, v\}, \end{aligned} \tag{19}$$

where i denotes the index of layers ranging from 1 to L, n denotes the index of the expert ranging from 1 to N. h_{tm}^{i} represents the cross-attention output in the i-th layer and $h_{t}^{1} = F_{t}$, $h_{m}^{1} = F_{m}$. Followed by [48], we use noisy gating to balance the load of different experts. Specifically, we use randomly initialized $W_{g} \in \mathbb{R}^{d_{model} \times N}$ and $W_{noise} \in \mathbb{R}^{d_{model} \times N}$ to map the weights logits of different experts, respectively. ϵ represents a matrix of Gaussian noise. H_{n}^{i} represents the logits of the n-th expert at the i-th layer. Top denotes the operation of keeping the top k largest values, with all other values being assigned as -∞.

4.2.2 Preemption State Estimation. When integrating three modalities, researchers often adopt a multi-stage fusion strategy [12, 18, 56, 70], where bi-modal fusions are performed first (stage 1), followed by a final integration of the fused modalities (stage 2). This paper hypothesizes that multimodal competition issues are present in each stage of the fusion process.

In Section 4.1.2, we have derived the Eq. 17 for estimating the modality preemption capability of modality m through evidence. Gradients are crucial in the learning process of neural networks, as they serve as the most direct representation of modal optimization conditions, and thus can also act as evidence of modal preemption speeds. We calculate the L2 norm of gradients from different encoders as evidence for estimating the preemption situation:

$$grad_{m} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} ||g_{i}^{m}||_{2}^{2}}, \tag{20}$$

where $m \in \{t, a, v, ta, tv\}$ represents the type of encoder, g_{i}^{m} denotes the grad of i-th parameter in m encoder, and n represents the number of parameter in m encoder. The $grad_{t}$, $grad_{a}$, $grad_{v}$ correspond to three unimodal encoders and $grad_{ta}$, $grad_{tv}$ correspond two multimodal encoders. We utilize these gradient norms to estimate the degree of causal preemption in stage 1 and stage 2, respectively.

Preemption State Estimation in Stage 1. In stage 1, we primarily compare the relationships among $grad_{t}$, $grad_{v}$, and $grad_{a}$. The Eq. 17 provides us with a method for causal estimation, from which we have expanded two simple parameter-free methods to estimate the current state: MAX and AVERAGE. The comparative results of the two strategies are in Section 5.8.

The MAX strategy assumes only one dominant modality in the current state. If the gradient norm of the current encoder is the maximum, it indicates that the current modality m_{i} has preempted the other modalities, marked as 1; otherwise, it is marked as 0:

$$P(u = \tilde{u}_{m_i} | e) = \begin{cases} 1, & \text{if } grad_{m_i} = \max(grad_{m_i}), \quad m_i \in \{t, a, v\}, \\ 0, & \text{others.} \end{cases} \tag{21}$$

The AVERAGE strategy assumes the existence of multiple dominant modalities at present. If the gradient norm of the current encoder is larger than the average of the other two encoders, it indicates that the current modality

m_i has preempted the other modalities, marked as 1; otherwise, it is marked as 0:

$$P(u=\tilde{u}_{m_i}|e)=\begin{cases}1, & \text{if } grad_{m_i} > \dfrac{grad_{m_j}+grad_{m_k}}{2}, \quad m_i,m_j,m_k\in\{t,a,v\}, i\neq j\neq k,\\ 0, & \text{others.}\end{cases} \tag{22}$$

Preemption State Estimation in Stage 2. In stage 2, we directly compare $grad_{tv}$ and $grad_{ta}$. The Cross-MoE with the larger gradient norm dominates the final fusion process, marked as 1; otherwise, it is marked as 0:

$$P(u=\tilde{u}_{m_i}|e)=\begin{cases}1 & \text{if } grad_{m_i}=\max(grad_{m_i}), \quad m_i\in\{ta,tv\},\\ 0 & \text{others.}\end{cases} \tag{23}$$

4.2.3 Adaptive Gradient Scaling. Due to complex data sampling and model learning states, modalities that exhibit good performance tend to dominate the training process, consequently causing varying degrees of under-optimized issues for other modalities. To address this problem, we dynamically adjust the training process of the well-performing modalities to mitigate the under-optimization issues of the other modalities using adaptive gradient scaling. We calculate the preemption ratio ρ_{m_i} using the following equation:

$$\rho_{m_i}=\begin{cases}\dfrac{2grad_{m_i}}{grad_{m_j}grad_{m_k}}, & \text{if } P(u=\tilde{u}_{m_i}|e)=1, \quad m_i,m_j,m_k\in\{t,a,v\}, i\neq j\neq k,\\ \dfrac{grad_{m_i}}{\max(grad_{m_i}grad_{m_j})}, & \text{elif } P(u=\tilde{u}_{m_i}|e)=1, \quad m_i,m_j\in\{ta,tv\}, i\neq j,\\ 0, & \text{others.}\end{cases} \tag{24}$$

Different stages of preemption are calculated separately, and a higher preemption rate indicates more severe modal competition in the current fusion stage. The assumption suggests that the preemption rate ρ_{m_i} should be negatively correlated with the strength of gradient scaling. That is, the stronger the optimization dominance of modality m_i over other modalities, the slower we would want the optimization of m_i to proceed. We define the following coefficient λ_{m_i} to dynamically regulate the training rates of the modal encoder corresponding to modality m_i:

$$\lambda_{m_i}=1-\tanh(\alpha\cdot\frac{\rho_{m_i}}{1+\rho_{m_i}}), \tag{25}$$

where α is a hyper-parameter to control the strength of gradient scaling. In practice, we use α and β to control the gradient scaling strength in the first stage and the second stage, respectively as discussed in Section 5.7. By aggregating the coefficient λ_{m_i} with the optimization method of Stochastic Gradient Descent (SGD) [1], we can achieve a balanced gradient update strategy in iteration t:

$$\theta^b_{t+1}=\theta^b_t-\eta\cdot\lambda_{m_i}\frac{1}{BS}\sum_{x\in B_t}\nabla_{\theta^b}\ell(x;\theta^b_t), \tag{26}$$

where BS denotes the batch size, B_t denotes the current batch of data of iteration t, b serves as an identifier for different encoders, θ^b_t represents the corresponding parameters in the encoder of modality m_i in iteration t. By dynamically scaling the gradients of the dominant modalities, we can alleviate the issue of modality competition.

4.2.4 Loss Function. We use F_{ta} and F_{tv} to denote the two output branches of the Cross-MoE $F_{ta}=h^{L+1}_{ta}$ and $F_{tv}=h^{L+1}_{tv}$, respectively. We concatenate these two vectors and then use a multilayer perceptron to produce the final sentiment logits. Following [13, 14, 20, 36], we use Mean Absolute Error as the loss function for multimodal sentiment analysis and cross-entropy as the loss function for multimodal humor detection. The overview of our algorithm in Algorithm 1:

$$\hat{y}=MLP([F_{ta};F_{tv}]), \quad Loss=\mathcal{L}_{task}(\hat{y},y). \tag{27}$$

Algorithm 1: Multimodal learning with AGS-SMoE.

Input : $D = \{(\mathbb{M}_t, \mathbb{M}_a, \mathbb{M}_v), Y\}$, γ, learning rate η_{main}, hyper-parameter α, β, N_{epochs}, initialized model f_θ
Output: Prediction $\hat{y}$ # *sentiment score*

for *each epoch* **do**
  Sample a mini-batch B_t from the dataset D;
  Process the batched data B_t using forward propagation;
  Calculate gradients for θ using backpropagation;
  Calculate the average L2 norm of the gradients for different encoders $grad_t$, $grad_a$, $grad_v$, $grad_{ta}$, and $grad_{tv}$ using Eq. 20;
  Calculate ρ_t, ρ_a, ρ_v, ρ_{ta}, and ρ_{tv} using Eq. 24;
  Calculate λ_t, λ_a, λ_v, λ_{ta}, and λ_{tv} using Eq. 25;
  Gradient scaling according to Eq. 26;
  Update model parameters using new gradients;

4.3 Actual Cause Guided Adaptive Gradient Scaling: Alleviating Multimodal Competition.

Traditional methods [11, 26, 46, 69] compare the inconsistency of additional unimodal classifier accuracies as an indicator of modality competition. Then they control the optimization speed of the dominant modality by modulating the gradient to keep all modalities under similar optimization states. However, blindly aligning the optimization of different modalities can also lead to a decrease in model accuracy for two main reasons [11]: 1. The inherent information content within modalities is unequal. Text has an innate intuitiveness for sentimental judgment [18, 20, 56], so the priority of the text modality should be taken into account. 2. There are inherent differences in the abilities of encoders to extract information from modalities. We can not force a perfectly trained encoder to align with a randomly initialized encoder under a similar optimization state, and vice versa.

Different from traditional methods [11, 26, 46, 69], AGS-SMoE replaces unimodal classifiers with the more stable and objective intrinsic gradient estimation of the actual cause, and aligns the optimization state with the actual cause. Ideally, the unimodal gradients, which represent the optimization speeds should be highly correlated with the actual causes of the unimodal on the fusion results. However, due to the multimodal competition, the correlation between gradients and actual cause is diminished. Because similar discriminative information may appear in different modalities simultaneously, once the fusion results have obtained sufficient discriminative information from the dominant modality preferentially, it will start to reject information from weaker modalities, leading to smaller gradients for the weaker modalities [21]. To address these issues, AGS-SMoE applies inverse weighted scaling to the gradients, limiting the dominant modality in the actual cause, to alleviate the information saturation of the fused modality. As a result, the weaker modality encoders will receive sufficient gradients and optimization speed. Due to the additional gradient calculations, similar to other gradient-based methods [11, 26, 46, 69], the training duration may be slightly extended.

AGS-SMoE can be considered as a special case of the Expectation-Maximization (EM) algorithm [41].

- In the Expectation step, AGS-SMoE estimates the actual cause u through current gradients $grad^c$, and due to modality competition, this is a negatively biased estimation for weaker modalities. Then, based on the $grad^c$ and u, we calculate the expectation likelihood of the ideal gradient $grad$. For the oppressed modality, the ideal gradient is the original gradient. For the dominant modality, the ideal gradient is the scaled gradient that can effectively balance its optimization and the subsequent optimization of other

modalities.

$$Q(grad|grad^c) = \mathbb{E}_{u|grad^c} L(grad|u, grad^c). \tag{28}$$

- In the Maximization step, AGS-SMoE adjusts the gradient of the dominant modality to maximize its expectation likelihood to the ideal gradient and obtain the gradient $grad^{c+1}$ adaptively for optimization. Through repeated iterations and alleviation of modality competition, the negatively biased estimation of the actual cause of weaker modalities towards the actual cause will gradually decrease in the Expectation step. Ultimately, all the gradients of modalities will align with the actual cause.

$$grad^{c+1} = \arg\max_{grad} Q(grad|grad^c) = AGS(grad^c). \tag{29}$$

5 EXPERIMENTS

5.1 Dataset

We conduct experiments on four publicly available datasets: CMU-MOSI [66], CMU-MOSEI [67], UR-FUNNY [14], and CH-SIMS [62]. The data splitting aligns the original datasets.

CMU-MOSI. The CMU-MOSI dataset is a comprehensive collection of 2199 opinionated video clips. Each clip is annotated with sentiment scores ranging from -3 to 3. It consists of 1,284 training samples, 229 validation samples, and 686 test samples, totaling 2,199 samples.

CMU-MOSEI. The CMU-MOSEI dataset is an expanded version of the CMU-MOSI. The data annotation standard is consistent with CMU-MOSI. It comprises 16,326 training samples, 1,871 validation samples, 4,659 test samples, and has a total of 22,856 samples.

UR-FUNNY. The UR-FUNNY dataset is the first dataset designed specifically for multimodal humor detection. Each video clip is annotated with a binary label, indicating whether it is humorous or non-humorous. It includes 7,614 training samples, 980 validation samples, 994 test samples, and sums up to 9,588 samples.

CH-SIMS. The CH-SIMS dataset is a MSA resource for the chinese language. Each clip is annotated with sentiment scores ranging from -1 to 1. It encompasses 1,368 training samples, 456 validation samples, 457 test samples, and has a total of 2,281 samples.

5.2 Evaluation Metrics

For CMU-MOSI and CMU-MOSEI, we use Mean Absolute Error (MAE) and Correlation Coefficient (Corr) to measure the accuracy of regression. We use ACC-2 and F1-score to measure the binary classification accuracy under two settings: positive/negative (P/N) and non-negative/negative (NN/N). In addition, we use ACC-7 to measure the accuracy of fine-grained sentiment classification within the range of -3 to 3.

For UR-FUNNY, we use ACC-2 to measure the binary classification accuracy.

For CH-SIMS, we evaluate the accuracy of regression using the Mean Absolute Error (MAE). For binary classification in negative/non-negative (N/NN) scenarios, we evaluate performance using the ACC-2 and F1-score metrics. Additionally, we measure the accuracy of fine-grained sentiment classification within the range of -1 to 1 using the ACC-5 metrics.

For the issue of modal competition, we report ACC-3 on CMU-MOSI, CMU-MOSEI, and CH-SIMS which predicts the three categories: negative, neutral, and positive.

5.3 Baselines

We compare AGS-SMoE with other baseline models that use the same preprocessing methods to validate its performance on the MSA task. These baselines include tensor-based fusion (**TFN** [64], **LMF** [33]), Transformer-based fusion (**MulT** [53], **MAG-BERT** [47], **BIMHA** [59], **TETFN** [56], **TGMN** [35], **TMBL** [19], **CRNet** [51]), and auxiliary task-guided fusion (**MISA** [15], **Self-MM** [63], **HyCon-BERT** [36], **MMIM** [13]).

Furthermore, we compared the AGS method with other modal competition mitigation approaches to verify the effectiveness of our proposed AGS in addressing modal competition issues. For a fair comparison, when conducting multimodal competition experiments, all models use the simplest multimodal concatenation method. Therefore, we do not use sparse MOE and will use AGS to represent our method. The baselines include traditional concat fusion, **G-Blending** [57], **OGM-GE** [46], **PMR** [8], **UME** and **UMT** [5].

5.4 Implement Details

Feature Extraction. For a fair comparison, we employ the same feature extraction method as HyCon-BERT [36], a method also utilized by the vast majority of baselines. Specifically, we use BERT [24] as the text encoder. For audio modality, we use COVERAP [4] to extract spectral features, speech polarity, harmonic parameters, etc. For visual modality, we use Facet [40] to extract action units, facial landmarks, and other expression data.

Hyperparameter Setting. We use Adam as the optimizer, with a learning rate of 1e-3 for all modules, and train with a fixed batch size of 64. The number of Cross-MoE layers is fixed at 9. The hyperparameters α and β that control the balance strength are tuned within the range of {0.001,0.005,0.01,0.05,0.1}. The hyperparameter γ that controls the end epoch of the gradient scaling is tuned within the range of {2,4,6,8} and the number of experts N is tuned within the range of {3,4,5}. We set K to 2, meaning up to K experts are activated. All experiments are conducted on a single Tesla V100 (32G). The sensitivity analysis of the hyperparameters is presented in Section 5.7.

5.5 Performance Results

Table 1. The comparative results of AGS-SMoE against other baselines on the CMU-MOSI and CMU-MOSEI datasets are as follows: For the metrics ACC-2 and F1, the values positioned to the left side of the "/" symbolize the outcomes in a non-negative/negative (NN/N) setting. Conversely, the figures on the right side denote the results in a positive/negative (P/N) setting. Metrics marked with an upward arrow ↑ signify that higher values correspond to superior performance, whereas those marked with a downward arrow ↓ indicate that lower values are indicative of better performance. The values that represented the state-of-the-art (SOTA) before this are underscored, and the current state-of-the-art (SOTA) values are presented in bold.

Models	CMU-MOSI					CMU-MOSEI				
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑
TFN[a] [64]	0.901	0.698	34.9	-/80.8	-/80.7	0.593	0.700	50.2	-/82.5	-/82.1
LMF[a] [33]	0.917	0.695	33.2	-/82.5	-/82.4	0.623	0.677	48.0	-/82.0	-/82.1
MulT[a] [53]	0.861	0.711	-	81.5/84.1	80.6/83.9	0.580	0.703	-	-/82.5	-/82.3
MAG-BERT[b] [47]	0.790	0.768	42.9	-/83.5	-/83.5	0.602	0.778	51.9	-/85.0	-/85.0
MISA[a] [15]	0.804	0.764	42.3	80.79/82.10	80.77/82.03	0.568	0.724	-	82.59/84.23	82.67/83.97
Self-MM[a] [63]	0.712	0.795	45.79	82.54/84.77	82.68/84.91	0.529	0.767	53.46	82.68/84.96	82.95/84.93
MMIM[a] [13]	0.700	0.800	46.65	84.14/86.06	84.00/85.98	0.526	0.772	54.24	82.24/85.97	82.66/85.94
HyCon-BERT[b] [36]	0.713	0.790	46.6	-/85.2	-/85.1	0.601	0.776	52.8	-/85.4	-/85.6
BIMHA[b] [59]	0.925	0.671	36.44	78.57/80.3	78.5/80.03	0.559	0.731	52.11	84.07/83.96	83.35/83.5
TETFN[b] [56]	0.717	0.800	-	84.05/86.10	83.83/86.07	0.551	0.748	-	84.25/85.18	84.18/85.27
TGMN[b] [35]	0.707	0.786	-	-/86.94	-/87.01	0.529	0.775	-	-/86.22	-/86.29
CRNet[b] [51]	0.712	0.797	**47.4**	-/86.4	-/86.4	0.541	0.771	53.8	-/86.2	-/86.1
TMBL[b] [19]	0.867	0.762	36.3	81.78/83.84	82.41/84.29	0.545	0.766	52.4	84.23/85.84	84.87/85.92
AGS-SMoE	**0.699**	**0.809**	46.06	**86.01/87.65**	**85.93/87.62**	**0.517**	**0.780**	**55.87**	**85.34/86.52**	**85.38/86.31**

[a]: results are from [13]; [b]: results are from corresponding original papers.

Table 2. The comparative outcomes between AGS-SMoE and other baseline models on CH-SIMS and UR-FUNNY are as follows. The marks in the table are consistent with those in Table 1.

Model	CH-SIMS				UR-FUNNY
	MAE↓	ACC-2↑	F1↑	ACC-5↑	ACC-2↑
TFN[a] [64]	0.432	78.38	78.62	39.30	68.57
LMF[a] [33]	0.441	77.77	77.88	40.53	67.53
MulT[a] [53]	0.453	78.56	79.66	37.94	70.55
MISA[a] [15]	-	-	-	-	70.61
Self-MM[a] [63]	0.425	80.04	80.44	41.53	-
TETFN[a] [56]	0.420	81.18	80.24	41.79	-
TGMN[b] [35]	-	81.18	81.43	-	-
CRNet[b] [51]	0.416	80.7	80.7	-	-
AGS-SMoE	**0.412**	**84.28**	**84.24**	**43.76**	**72.13**

[a]: results are from [37] and its corresponding github page; [b]: results are from corresponding original papers.

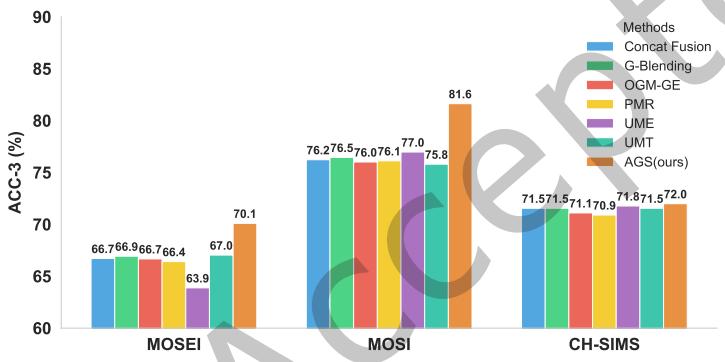


Fig. 6. The comparison results with multimodal methods that focus on multimodal competition issues.

The comparative results of AGS-SMoE with other baseline models on the CMU-MOSI and CMU-MOSEI datasets are illustrated in Table 1. Additionally, the comparative results on the CH-SIMS and UR-FUNNY datasets are presented in Table 2. The experimental results concerning the multimodal competition issue are depicted in Fig. 6. AGS-SMoE has achieved state-of-the-art (SOTA) status in nearly all of the measured metrics.

On the CMU-MOSI dataset, AGS-SMoE achieves state-of-the-art results in all metrics except for the ACC-7 metric. Specifically, the ACC-2 and F1-score improve by nearly 2% under the non-negative/negative setting and 0.6% under the positive/negative setting. There are also improvements in regression accuracy metrics such as MAE and Corr. The experiments demonstrate that even without employing complex training processes like contrastive learning or other auxiliary tasks, we can alleviate the modality competition and enhance the model performance through gradient scaling and MoEs.

On the CMU-MOSEI dataset, AGS-SMoE achieves state-of-the-art results across all metrics, with improvements of 0.5~1% in ACC-2 and 0.3% in F1 accuracy, and also showed significant improvements in regression accuracy. Notably, AGS-SMoE achieved a 1.63% improvement in fine-grained classification ACC-7 compared to the previous state-of-the-art. The baselines are often affected by modal competition, making it difficult to leverage the advantages of multimodal. These results demonstrate that AGS-SMoE can handle modality competition in more complex scenarios and improve the model performance.

On the CH-SIMS and UR-FUNNY datasets, AGS-SMoE achieves state-of-the-art results across all metrics. Notably, AGS-SMoE achieved improvements of 3.1% in ACC-2 and 2.81% in F1 compared to the previous best values and improved ACC-5 by 1.97% on CH-SIMS. Furthermore, AGS-SMoE achieved an improvement of 0.45% in ACC-2 on UR-FUNNY. These results validate the general applicability of AGS-SMoE across different scenarios and language environments.

When focusing on the issue of multimodal competition, AGS-SMoE can also achieve the optimal values among comparable methods. Specifically, compared to the previous best ACC-3 scores, AGS has improved by 3.1% on CMU-MOSI, 4.6% on CMU-MOSEI, and 0.2% on CH-SIMS. The experimental results confirm that our proposed AGS method can more adaptively adjust gradients to mitigate the modal competition issue compared to other methods. This demonstrates that it is feasible to assess the modal preemption state in the dynamic training process by evaluating the contribution of modal pairs to the fusion result.

Our proposed AGS-SMoE achieves state-of-the-art results across four public datasets and outperforms baselines on the multimodal competition problem, demonstrating that our method can alleviate modality competition, promote modal fusion, and improve model performance.

5.6 Ablation Study

Table 3. The ablation results on the CMU-MOSI, CMU-MOSEI datasets.

Models	CMU-MOSI					CMU-MOSEI				
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑
AGS-SMoE	0.699	**0.809**	46.06	**86.01/87.65**	**85.93/87.62**	**0.517**	**0.780**	**55.87**	**85.34/86.52**	**85.38/86.31**
w/o taMoE	0.714	0.800	45.48	84.26/86.59	84.13/86.53	0.522	0.776	54.53	84.48/85.86	84.56/85.64
w/o tvMoE	0.700	0.800	47.23	84.55/86.59	84.36/86.47	0.524	0.776	54.45	84.80/85.88	84.81/85.63
w/o allMoE	0.717	0.800	45.91	84.26/86.13	84.21/86.13	0.536	0.766	54.07	84.31/85.66	84.38/85.44
w/o Grad_t	0.697	0.803	47.08	84.55/86.13	84.49/86.11	0.532	0.761	54.23	84.16/85.91	84.19/85.63
w/o Grad_v	0.693	0.808	**48.10**	84.69/87.03	84.48/86.91	0.531	0.775	54.36	84.55/86.30	84.72/86.17
w/o Grad_a	**0.689**	0.807	46.79	84.40/86.89	84.22/86.79	0.531	0.777	54.26	84.33/85.99	84.47/85.82
w/o Grad_{stage1}	0.705	0.805	46.64	83.38/85.06	83.35/85.07	0.529	0.767	54.13	83.86/85.66	84.08/85.55
w/o Grad_{stage2}	0.702	0.804	45.62	84.55/86.28	84.45/86.24	0.536	0.763	54.28	84.22/85.66	84.30/85.44
w/o Grad_{all}	0.722	0.799	45.34	83.97/85.82	83.84/85.75	0.532	0.762	54.97	83.11/84.89	83.15/84.58

We conduct ablation studies on the MoE module (the top part of Table 3) and the gradient scaling (the bottom part of Table 3) separately to verify their roles in MSA. The results are listed in Table 3.

The effect of the sparse mixture-of-experts module (the top part of Table 3). To validate the role of the MoE module, we individually replaced the text-audio branch (denoted as w/o taMoE), the text-video branch (denoted as w/o tvMoE), and all branches (denoted as allMoE) with the traditional transformer to observe the differences in their effects. We find that the MoE module plays a significant role in the performance on both datasets and removing any branch will lead to a decrease in accuracy. On the CMU-MOSI dataset, removing the taMoE branch results in a greater loss of accuracy than removing the tvMoE branch, while there is no significant difference on the CMU-MOSEI dataset. This may imply that text and audio are more likely to compete on CMU-MOSI, requiring appropriate sparsity to alleviate their competitive relationship. Removing all MoE branches results in the worst performance on both datasets, confirming that using the MoE module in fusion can alleviate modality competition and improve model accuracy.

The effect of the gradient scaling operations (the bottom part of Table 3). To validate the role of gradient scalings, we individually removed the audio modality grad scaling (denoted as Grad_a), visual modality grad scaling (denoted as Grad_v), text modality grad scaling (denoted as Grad_t), all unimodal grad scalings (denoted as Grad_{stage1}), all bimodal grad scalings (denoted as Grad_{stage2}), and all grad scalings (denoted as Grad_{all}), and

observed the differences in their effects. By comparing the gradient scalings of single modalities, we find that removing the gradient scaling for text leads to the greatest decrease in accuracy, suggesting that text has a higher probability of suppressing the learning of other unimodal encoders. Furthermore, we find that, compared to the visual modality, the audio modality has a greater tendency to suppress other modalities, which may be due to the audio modality contains more sentimental information [11, 46]. In addition, we find that removing the gradient scaling of any modality results in a performance decrease, indicating that due to data distribution and the dynamic learning process, there is mutual competition and suppression among different modalities. This also confirms that gradient scaling is an effective means to alleviate modality competition. When all single-modality gradient scalings are removed, the accuracy drops the most significantly. Lastly, we find that modality competition exists in modality fusion at different stages, and gradient scaling at different stages is very necessary for improving model accuracy, which also verifies our previous conjecture. The gradient scaling at the first level can be more effective compared to the gradient scaling at the second level.

Through ablation experiments, we confirm that the various components in our proposed AGS-SMoE can effectively alleviate modality competition, promote modality fusion, and enhance the model performance across different datasets.

5.7 Sensitivity and Quantitative Analysis

In this section, we conduct a sensitivity analysis on hyperparameters, a quantitative plug-in strategy analysis, and a quantitative analysis of different gradient scaling strategies.

Table 4. Sensitive analysis of α.

α	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
0.001	0.701	84.55/86.59	0.521	83.86/86.41
0.005	**0.692**	84.26/86.28	0.528	83.22/85.94
0.01	0.699	**86.01**/**87.65**	0.524	84.59/**86.74**
0.05	0.694	84.69/86.28	**0.517**	**85.34**/86.52
0.1	0.696	84.84/86.59	0.527	84.20/86.10

Table 5. Sensitive analysis of β.

β	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
0.001	0.717	84.26/86.59	0.521	83.86/86.41
0.005	0.717	84.69/86.42	**0.517**	**85.34**/**86.52**
0.01	**0.699**	**86.01**/**87.65**	0.528	84.83/85.80
0.05	0.732	84.99/87.04	0.529	85.30/86.32
0.1	0.707	84.99/86.13	0.536	84.12/86.41

Sensitivity analysis on the α and β. In addition, we conduct sensitivity tests on the hyperparameters α and β, which control the intensity of gradient scaling in the unimodal and bimodal process, respectively. The results are shown in Table 4 and Table 5. We find that on the CMU-MOSI dataset, the optimal combination of α and β is 0.005 and 0.5, while on the CMU-MOSEI it is 0.01 and 0.005. This indicates that the modal competition issue is prevalent at different fusion stages across different datasets, and the intensity of modal competition also varies. This also confirms that an appropriate intensity of gradient scaling is beneficial in alleviating modal competition.

Table 6. Sensitive analysis of the number of experts in MoE.

experts	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
3	0.699	**86.01**/**87.65**	**0.517**	**85.34**/86.52
4	0.729	85.42/87.04	0.524	84.42/**86.53**
5	**0.696**	84.84/86.59	0.529	85.30/86.32

Table 7. Sensitive analysis of end epoch.

end epoch	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
2	**0.696**	84.84/86.59	**0.517**	85.34/86.52
4	0.697	84.69/86.74	0.530	**85.62**/86.42
6	0.699	**86.01**/**87.65**	0.524	84.42/**86.53**
8	0.697	84.69/86.74	0.521	83.86/86.41

Sensitivity analysis on the number of experts. We first conduct a sensitivity analysis on the number of experts in the MoE, and the results are shown in Table 6. We find that the optimal number of experts is 3. When the number of Experts increases, the model does not adapt to complex datasets as expected but encounters

difficulties during the optimization process, which may be due to the larger parameter space of the model. When we set top-2 gate activation, the probability of each expert receiving a reduced number of tokens also leads to some experts being underfitted. The experimental results indicate that an appropriate number of experts can sparsify modal fusion, while too many experts can hinder model optimization and reduce accuracy.

Sensitivity analysis on the end epoch. We conduct a sensitivity test on the end epoch for gradient scaling, with the results shown in Table 7. To alleviate the issue of modal competition, we have introduced gradient scaling, but this has made the model's optimization perpetually suboptimal. Therefore, we need to stop the gradient scaling at an appropriate epoch, allowing the model to overcome the modal competition early in training and gradually converge to the global optimum. We also observe that the optimal end epoch differs due to varying convergence conditions on different datasets. On the CMU-MOSI dataset, the optimal end epoch is 6, while on the CMU-MOSEI dataset, it is 2.

Table 8. The plug-in strategy of AGS and average training duration (sec) across five random seeds.

Models	CMU-MOSI			CMU-MOSEI		
	ACC-2↑	F1↑	time(sec)	ACC-2↑	F1↑	time(sec)
MISA	80.79/82.10	80.77/82.03	413.25	82.59/84.23	82.67/83.97	1339.98
+AGS(Ours)	**83.67/85.52**	**83.59/85.49**	423.59	**83.99/85.61**	**84.07/85.37**	1365.62
Δ_{AGS}	↑2.88/↑3.42	↑2.82/↑3.46	10.34	↑1.40/↑1.38	↑1.40/↑1.40	25.64
Self-MM	82.54/84.77	82.68/84.91	285.13	82.68/84.96	82.95/84.93	1132.78
+AGS(Ours)	**85.28/87.35**	**85.15/87.28**	293.70	**84.67/86.02**	**84.75/85.82**	1154.63
Δ_{AGS}	↑2.74/↑2.58	↑2.47/↑2.37	8.57	↑1.99/↑1.06	↑1.80/↑0.89	21.85
MMIM	84.14/86.06	84.00/85.98	605.38	82.24/85.97	82.66/85.94	2159.03
+AGS(Ours)	**85.28/86.89**	**85.25/86.90**	614.18	**84.14/86.38**	**84.27/86.18**	2187.94
Δ_{AGS}	↑1.14/↑0.83	↑1.25/↑0.92	9.80	↑1.90/↑0.41	↑1.61/↑0.24	28.91

Analysis on the plug-and-play capability of AGS. To verify that our proposed gradient scaling strategy can be flexibly and seamlessly integrated into other models, we have combined our gradient scaling strategy with three open-source MSA models (MISA [15], Self-MM [63], and MMIM [13]) and conducted relevant experiments on the CMU-MOSI and CMU-MOSEI datasets as listed in Table 8. The AGS we proposed significantly enhances the performance of existing methods on multimodal sentiment analysis tasks and is adept at adapting to multi-task guided multimodal learning. Experimental results have shown that the AGS is a flexible plug-and-play module that can adapt to various fusion methods and models. It is worth noting that the training duration does indeed show a slight increase, which is consistent with the theoretical analysis in Section 4.3. However, in practical comparisons with the baseline model, the slowdown is negligible.

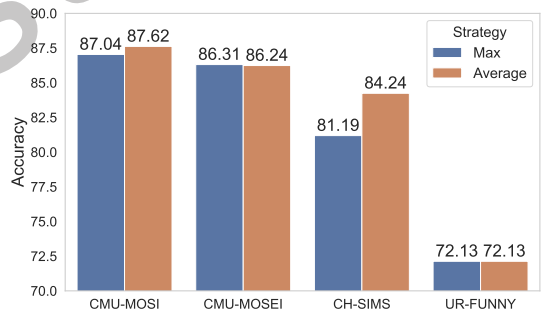
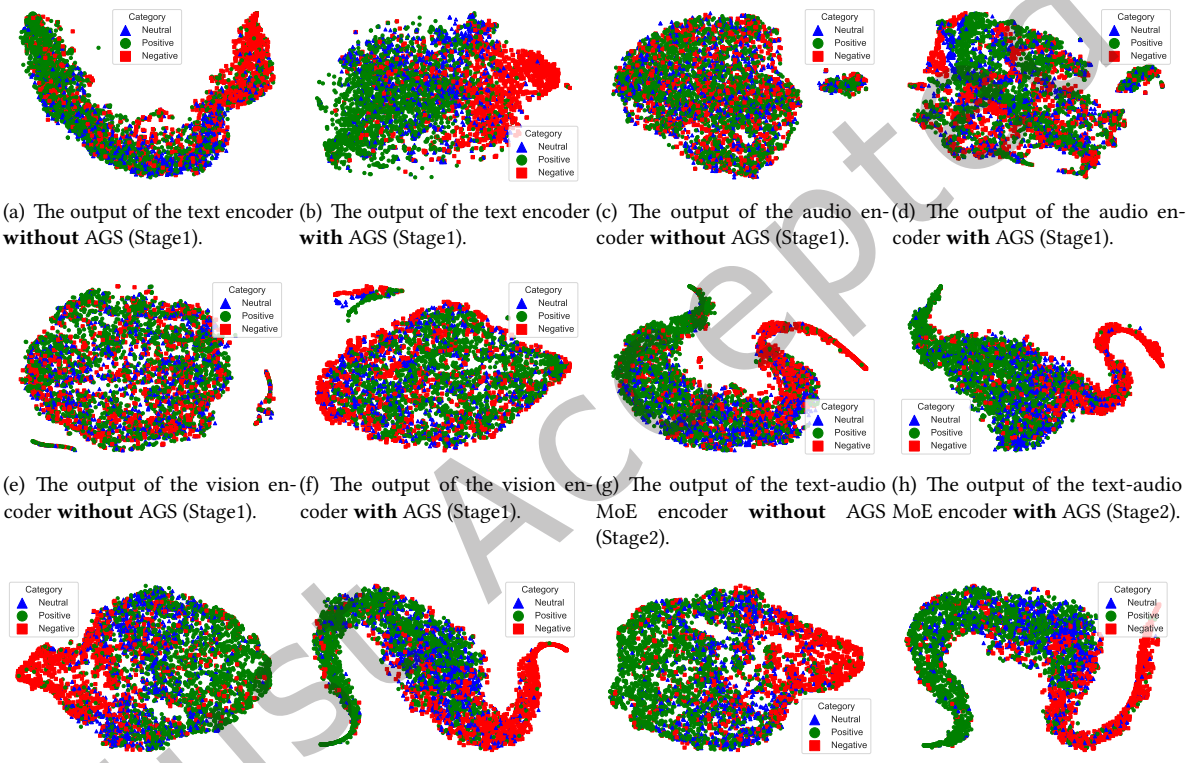


Fig. 7. Comparison of different gradient scaling strategies on four datasets.

Analysis of the different gradient scaling strategies. We compare the performance of different gradient scaling strategies on four datasets (the f1-score under positive/negative settings for CMU-MOSI, CMU-MOSEI

and CH-SIMS, ACC-2 for UR-FUNNY) as shown in Fig. 7. "Max" indicates that we only adjust the gradients for the encoder corresponding to the maximum gradient norm, while "Average" indicates that we adjust the gradients by comparing the mean gradient norms among the three modalities, as described in Section 4.2.2. We find that the Max gradient scaling strategy performed better than the Average gradient scaling strategy on the CMU-MOSEI dataset, while the Average gradient scaling strategy outperformed the Max gradient scaling strategy on the CMU-MOSI and CH-SIMS datasets, and there was no significant difference between the two gradient scaling strategies on UR-FUNNY. These two strategies correspond to different assumptions and should be tailored according to the dataset.

5.8 Visualization



(a) The output of the text encoder **without** AGS (Stage1).
(b) The output of the text encoder **with** AGS (Stage1).
(c) The output of the audio encoder **without** AGS (Stage1).
(d) The output of the audio encoder **with** AGS (Stage1).
(e) The output of the vision encoder **without** AGS (Stage1).
(f) The output of the vision encoder **with** AGS (Stage1).
(g) The output of the text-audio MoE encoder **without** AGS (Stage2).
(h) The output of the text-audio MoE encoder **with** AGS (Stage2).
(i) The output of the text-vision MoE encoder **without** AGS (Stage2).
(j) The output of the text-vision MoE encoder **with** AGS (Stage2).
(k) The final fusion result **without** AGS (Stage3).
(l) The final fusion result **with** AGS (Stage3).

Fig. 8. The t-SNE visualization analysis results of the outputs from different encoders at various fusion stages on the CMU-MOSEI test set. Even columns use the AGS method, odd columns do not.

To further investigate the role of AGS at different fusion stages and in different encoders, we visualized the result of AGS-SMoE and its variant (without AGS) on the CMU-MOSEI dataset using t-SNE [54] as shown in Fig. 8.

In unimodal feature extraction **(Stage1)**, by comparing Fig. 8(a) and Fig. 8(b), Fig. 8(c) and Fig. 8(d), and Fig. 8(e) and Fig. 8(f), it can be observed that without the use of gradient scaling strategies, the representation learning of

both the audio and visual modalities is poor, as they are dominated by the textual modality. After employing the gradient scaling strategy, the distinctiveness of all modalities in the feature space has been enhanced. Even though the text modality plays a dominant role in the vast majority of cases and is only suppressed in a few iterations, our gradient scaling strategy can still promote the learning of text representations. This confirms that the gradient scaling strategy can enhance the learning of representations for all single modalities. In multimodal fusion **(Stage2)**, by comparing Fig. 8(g) and Fig. 8(h), Fig. 8(i) and Fig. 8(j). Because the information of the text modality is integrated, the output results of text-audio and text-vision have both been significantly enhanced. However, without using AGS, the decision boundary for sentiment remains unclear. Using gradient scaling in stage two can also promote the representation of the fused modalities. In addition, by comparing Fig. 8(k) and Fig. 8(l), we find that the final tri-modal representation **(Stage3)** has also been optimized, showing a more distinct sentimental differentiation. In summary, compared to the baseline without AGS, the method using AGS has clearer decision boundaries, improved feature alignment and fusion result, and more structured trimodal representation. Our proposed AGS can alleviate the modal competition issues and promote the representation learning of different stages.

6 CONCLUSION

In this paper, we propose an effective learning method called Adaptive Gradient Scaling with sparse Mixture-of-Experts (AGS-SMoE) to alleviate the problem of modal competition. The core of AGS-SMoE is modal preemption state estimation and adaptive gradient scaling. The AGS-SMoE achieved the best results on the four datasets in multimodal sentiment analysis. In addition, AGS-SMoE significantly enhances representation learning at different fusion stages using a parameter-free approach. We hope that this work will provide a fresh perspective on modal competition and inspire better multimodal model designs. The limitation of our method is that it may slightly slow down the training of the model. Future work can focus on directly integrating multimodal balancing mechanisms into the fusion module without any additional guidance.

REFERENCES

[1] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Springer, 177–186.

[2] Martin Bunzl. 1980. Causal preemption and counterfactuals. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 37, 2 (1980), 115–124.

[3] Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *Comput. Surveys* 55, 13s (2023), 1–38.

[4] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[5] 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*. PMLR, 8632–8656.

[6] Douglas Ehring. 1984. Probabilistic causality and preemption. *The British Journal for the Philosophy of Science* 35, 1 (1984), 55–57.

[7] Douglas Ehring. 1990. Preemption, direct causation, and identity. *Synthese* (1990), 55–70.

[8] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20029–20038.

[9] Wenjun Feng, Xin Wang, Donglin Cao, and Dazhen Lin. 2024. An autoencoder-based self-supervised learning for multimodal sentiment analysis. *Information Sciences* 675 (2024), 120682.

[10] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion* 91 (2023), 424–444.

[11] Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. 2024. Classifier-guided Gradient Modulation for Enhanced Multimodal Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[12] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*.

6–15.
[13] Wei Han, Hui Chen, and Soujanya Poria. 2021. 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*. 9180–9192.
[14] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2046–2056.
[15] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 1122–1131.
[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[17] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 7837–7851.
[18] Changqin Huang, Junling Zhang, Xuemei Wu, Yi Wang, Ming Li, and Xiaodi Huang. 2023. TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis. *Knowledge-Based Systems* 269 (2023), 110502.
[19] Jiehui Huang, Jun Zhou, Zhenchao Tang, Jiaying Lin, and Calvin Yu-Chian Chen. 2024. TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowledge-Based Systems* 285 (2024), 111346.
[20] Qionghao Huang, Jili Chen, Changqin Huang, Xiaodi Huang, and Yi Wang. 2024. Text-centered cross-sample fusion network for multimodal sentiment analysis. *Multimedia Systems* 30, 4 (2024), 228.
[21] 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*. PMLR, 9226–9259.
[22] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. 2024. C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16006–16015.
[23] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
[24] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
[25] Kyeonghun Kim and Sanghyun Park. 2023. AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis. *Information Fusion* 92 (2023), 37–45.
[26] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. 2023. Boosting Multi-modal Model Performance with Adaptive Gradient Modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22214–22224.
[27] Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. 2024. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *arXiv preprint arXiv:2405.05949* (2024).
[28] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2024. Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts. *arXiv preprint arXiv:2405.11273* (2024).
[29] Zuhe Li, Qingbing Guo, Yushan Pan, Weiping Ding, Jun Yu, Yazhou Zhang, Weihua Liu, Haoran Chen, Hao Wang, and Ying Xie. 2023. Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis. *Information Fusion* 99 (2023), 101891.
[30] Zuhe Li, Zhenwei Huang, Yushan Pan, Jun Yu, Weihua Liu, Haoran Chen, Yiming Luo, Di Wu, and Hao Wang. 2024. Hierarchical denoising representation disentanglement and dual-channel cross-modal-context interaction for multimodal sentiment analysis. *Expert Systems with Applications* 252 (2024), 124236.
[31] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947* (2024).
[32] Dahuang Liu, Zhenguo Yang, and Zhiwei Guo. 2024. Progressive Fusion Network with Mixture of Experts for Multimodal Sentiment Analysis. In *2024 16th International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, 150–157.
[33] Zhun Liu and Ying Shen. 2018. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.
[34] Ziyu Liu, Tao Yang, Wen Chen, Jiangchuan Chen, Qinru Li, and Jun Zhang. 2024. Sentiment analysis of social media comments based on multimodal attention fusion network. *Applied Soft Computing* (2024), 112011.
[35] Yuanyi Luo, Rui Wu, Jiafeng Liu, and Xianglong Tang. 2023. A text guided multi-task learning network for multimodal sentiment analysis. *Neurocomputing* 560 (2023), 126836.
[36] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing* (2022).
[37] Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-SENA: An Integrated Platform for Multimodal Sentiment Analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 204–213.

[38] Saeed Masoudnia and Reza Ebrahimpour. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review* 42 (2014), 275–293.
[39] Peter Menzies. 1996. Probabilistic causation and the pre-emption problem. *Mind* 105, 417 (1996), 85–117.
[40] Thomas B Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal. 2011. *Visual analysis of humans*. Springer.
[41] Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine* 13, 6 (1996), 47–60.
[42] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems* 35 (2022), 9564–9576.
[43] MZ Naser. 2022. Causality, causal discovery, and causal inference in structural engineering. *arXiv preprint arXiv:2204.01543* (2022).
[44] Judea Pearl. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.
[45] Judea Pearl. 2009. *Causality*. Cambridge university press.
[46] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8238–8247.
[47] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
[48] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).
[49] Li Shen, Anke Tang, Enneng Yang, Guibing Guo, Yong Luo, Lefei Zhang, Xiaochun Cao, Bo Du, and Dacheng Tao. 2024. Efficient and effective weight-ensembling mixture of experts for multi-task model merging. *arXiv preprint arXiv:2410.21804* (2024).
[50] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. 2023. Scaling Vision-Language Models with Sparse Mixture of Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 11329–11344.
[51] Hang Shi, Yuanyuan Pu, Zhengpeng Zhao, Jian Huang, Dongming Zhou, Dan Xu, and Jinde Cao. 2024. Co-space Representation Interaction Network for multimodal sentiment analysis. *Knowledge-Based Systems* 283 (2024), 111149.
[52] Lorenzo Stacchio, Alessia Angeli, Giuseppe Lisanti, Daniela Calanca, and Gustavo Marfia. 2022. Toward a holistic approach to the socio-historical analysis of vernacular photos. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 3s (2022), 1–23.
[53] 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 conference. Association for computational linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.
[54] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[56] Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. 2023. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition* 136 (2023), 109259.
[57] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12695–12705.
[58] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. 2020. On modality bias in the tvqa dataset. *arXiv preprint arXiv:2012.10210* (2020).
[59] Ting Wu, Junjie Peng, Wenqiang Zhang, Huiran Zhang, Shuhua Tan, Fen Yi, Chuanshuai Ma, and Yansong Huang. 2022. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowledge-Based Systems* 235 (2022), 107676.
[60] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. 2024. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739* (2024).
[61] Shu Yang, Muhammad Asif Ali, Cheng-Long Wang, Lijie Hu, and Di Wang. 2024. MoRAL: MoE Augmented LoRA for LLMs' Lifelong Learning. *arXiv preprint arXiv:2402.11260* (2024).
[62] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3718–3727.
[63] 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 *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 10790–10797.
[64] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114.
[65] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
[66] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.

[67] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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)*. 2236–2246.

[68] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27456–27466.

[69] Yupei Zhang, Xiaofei Wang, Fangliangzi Meng, Jin Tang, and Chao Li. 2024. Knowledge-Driven Subspace Fusion and Gradient Coordination for Multi-modal Learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 263–273.

[70] Chuanbo Zhu, Min Chen, Sheng Zhang, Chao Sun, Han Liang, Yifan Liu, and Jincai Chen. 2023. SKEAFN: sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis. *Information Fusion* 100 (2023), 101958.

[71] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion* 95 (2023), 306–325.