Customized-Allocation MoE: Reliable Dynamic Combination via the Attraction-Repulsion for Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Multimodal Sentiment Analysis (MSA) aims to infer sentiment by jointly leveraging textual, visual, and acoustic modalities. However, a core challenge remains: *How to dynamically identify and leverage sample-level dependency preferences among different modality combinations?* To address this, we decompose the challenge into two subproblems: combination matching and fusion validation. Correspondingly, we propose the Customized-Allocation Mixture-of-Experts (CA-MoE), a novel framework that consists of two core complementary components that enable dynamic and sample-level modality routing within the MoE architecture. First, Affinity-guided Customized Modality Allocation (ACMA) acts as a distributor and leverages Geometry-Gradient Affinity (G^2-Affinity) to guide an attraction-repulsion routing mechanism for customized allocation of modality combinations. After expert fusion, we then design Reliability-Aware Expert Selection (RAES) to jointly consider sentiment angular-prototype proximity and competitive magnitude intensity of representation. This yields a reliability selection matrix that weights over experts for the final sentiment prediction. Extensive experiments on three benchmark MSA datasets demonstrate that CA-MoE achieves significant or competitive performance over state-of-the-art methods.

1 Introduction

Human emotional expression is inherently multimodal, interweaving spoken language, facial expressions, and vocal cues. Therefore, unimodal approaches inevitably lose complementary affective signals, leading to incomplete sentiment understanding. (Baltruaitis et al., 2019). Multimodal Sentiment Analysis (MSA) aims to predict sentiment by jointly modeling text, vision, and audio, to reduce modality-specific bias that arises when any single modality is used alone. (Poria et al., 2017). In contrast to unimodal approaches, MSA has therefore become a cornerstone task in affective computing (Yu et al., 2024; Yang et al., 2025a; Zhao et al., 2025a; Ahuja et al., 2025).

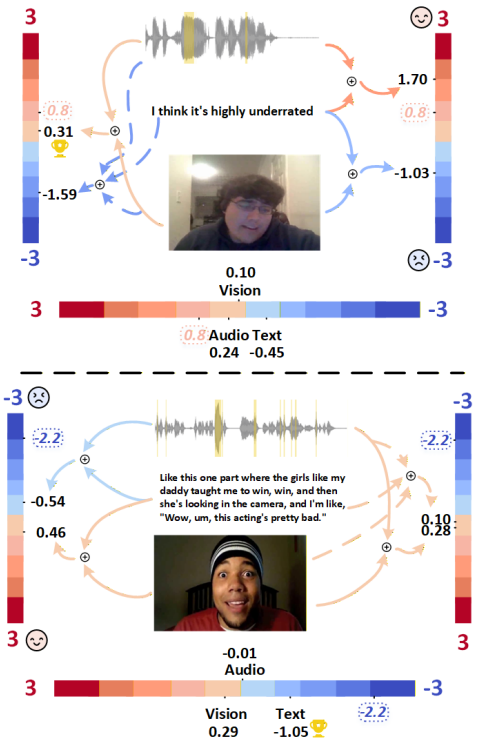


Figure 1: Illustrative example of two samples in dataset of CMU-MOSI (Zadeh et al., 2016) with widely-recognized MISA (Hazarika et al., 2020), which quantitatively demonstrates that predictions obtained from different modality combinations.

Despite strong progress, most MSA models still rely on a fixed modality interaction pattern. A common line of work either treats all modalities as equally informative or assumes a dominant anchor modality (often text) to guide fusion. However, in real-world scenarios, an important challenge is that modality interaction pattern often shows great dynamics. Specifically, the most reliable modality (or modality combination) can vary substantially

across samples, and cross-modal conflicts can be hierarchical and complicated. For example, the combination of a smiling expression and positive words delivers the positive signal, whereas audio represents sarcasm (Bao et al., 2025). In this case, multimodal interaction patterns with equal treatment or over-reliance on unimodal are both failing to resolve cross-modal polarity conflicts.

To further explore the existence and significance of this challenge, we conducted a pilot study. As shown in Fig. 1, three key findings emerge: (1) *Trimodal fusion is not universally optimal*: predictions from full-modality fusion can deviate substantially from ground truth, even exhibiting opposite polarity (e.g., -1.59 vs. 0.8). (2) *No single modality consistently dominates*: in the upper sample, the audio-visual combination (0.31) outperforms all text-involved combinations, contradicting the common assumption of text dominance. (3) *Optimal combinations are sample-dependent*: the upper sample favors bimodal fusion while the lower benefits from unimodal input. Moreover, within a single sample, different combinations can yield conflicting polarities (-0.54 vs. 0.46), making indiscriminate fusion unreliable.

These observations motivate a central question: **how can we dynamically identify and leverage sample-level preferences over modality combinations?** We decompose this question into two subproblems. ① **Combination matching**: for a given sample, which modalities should be fused together and which should be kept separate? ② **Fusion validation**: among the activated unimodal/bimodal/trimodal fusions, which fused representations are reliable for final prediction?

To address the above question, we propose **Customized-Allocation Mixture-of-Experts (CA-MoE)**, a novel framework that enables dynamic, sample-adaptive modality routing within the MoE architecture. Specifically, CA-MoE consists of two core complementary components:① **Affinity-guided Customized Modality Allocation (ACMA)** acts as a *distributor* and solves combination matching by routing modality combinations through an attraction–repulsion mechanism, guided by Geometry–Gradient Affinity (G^2-Affinity), which jointly reflects semantic alignment and optimization compatibility. ② **Reliability-Aware Expert Selection (RAES)** is proposed as a *validator*, which comprehensively evaluates the reliability from the sentiment angular-prototype proximity and competitive magnitude intensity of representation, producing a reliability weighting over experts for the final decision.

In summary, CA-MoE explicitly and dynamically models sample-level dependency preferences over modality combinations and prioritizes reliable fused representations to improve sentiment prediction. Our main contributions are three-folds:

- We identify and analyze an underexplored yet challenging problem in MSA: How to dynamically identify and leverage sample-level dependency preference among different modality combinations?
- We propose CA-MoE featuring ACMA for identifying and routing sample-varied modal combinations to expert and RAES for validating the reliability after-fused expert representation to encourage most reliable experts contribute to the final sentiment decision.
- Extensive experiments on three widely-used MSA datasets demonstrate state-of-the-art performance. Comprehensive ablations validate the contribution of each proposed component.

2 Related Work

Most existing MSA approaches based on assumptions on modality priors can be categorized into two groups: equally informative assignment methods and dominant modality assignment methods.

Equally informative assignment methods. Previous works in this group treat all modalities' information as a priori and pursue performance gains through capturing richer semantic understanding by designing high-capacity encoders (He et al., 2025; Zhuang et al., 2024) or optimizing representation strategy (Yu et al., 2021a; Hazarika et al., 2020). Moreover, many studies also focus on cross-modal interaction mechanisms. Early approaches (Zadeh et al., 2017; Liu et al., 2018) typically advocate for explicit interaction mechanisms for cross-modal dynamic dependencies. Compared with their static fusion paradigm, recent advances have turned to cross-modal Transformer architectures, which leverage cross-attention mechanisms for sample-level interactions in a dynamic manner (Huang et al., 2024; Xu et al., 2025; Gong et al., 2026; Zhao et al., 2025b).

Dominant modality assignment methods. These methods emphasize semantic guidance from a dominant modality, usually text (Yang et al.,

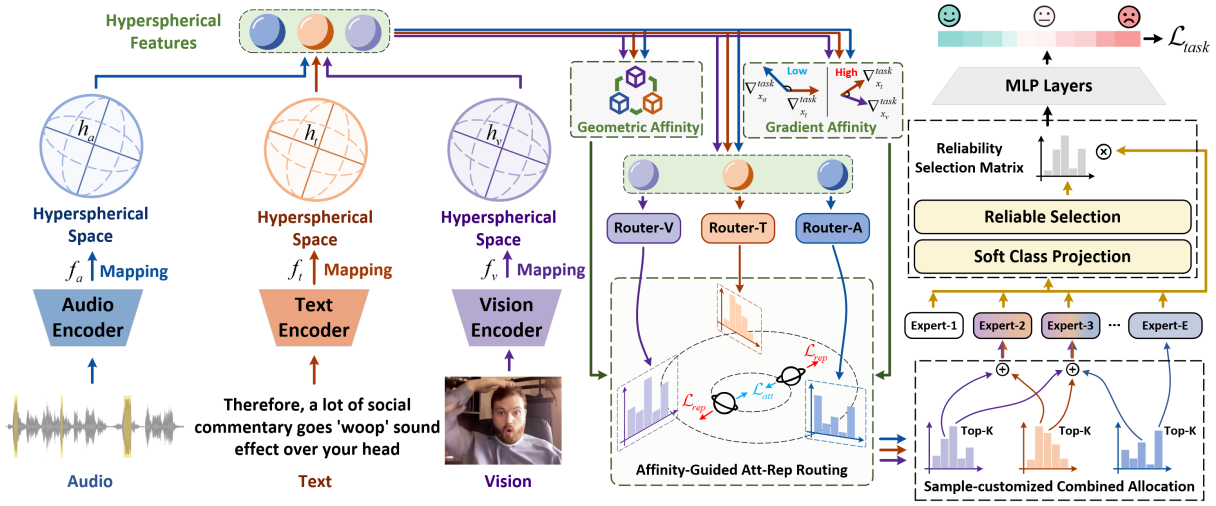


Figure 2: The overall architecture of **CA-MoE**. The blue, orange, and purple represent the audio, text, and vision modalities. Especially, we utilize colorless to indicate that there is no modal index to this expert. Simultaneously, various combinations of colors and quantities represent different modal combinations. For example, Expert-2 with two different color means the bimodal combination with text and vision.

2025b). They treat the dominant modality as an anchor and align auxiliary modalities through text-centric interaction mechanisms (Wang et al., 2025; Wu et al., 2025a; Yang et al., 2025c). However, recent studies have noted the issue of semantic noise introduced by text modality as a fixed prior. Some researches enable the model to dynamically select the dominant modality and adjust the contribution of each modality through additional unimodal modal supervision (Feng et al., 2024), sample gradient (Li et al., 2024), and routing mechanism (Gao et al., 2024; Fang et al., 2025).

Despite the promising results achieved for MSA, the above approaches ignore the possibility of modality combination as dominance, and these methods share an underlying limitation: their modality interaction patterns are usually fixed multimodal a priori and applied uniformly across all samples. However, in real-world applications, such an interaction pattern is often highly dynamic and even contains complicated, hierarchical intermodal conflicts. To solve this, we propose a dynamic modalities allocation approach to flexibly customize modal combinations and interaction. Furthermore, we prioritize reliability after-fused representation, thus improving predictions.

3 Methodology

3.1 Overall Architecture

As shown in Fig 2, which shows the overall workflow of CA-MoE. Specifically, we first extracts unimodal representations with corresponding encoders. The unimodal representations are mapped into hyperspherical space to obtain hyperspherical features for emotional polarity expression (Sec3.3). Then, Affinity-guided Customized Modality Allocation (ACMA) leverages Geometry-Gradient Affinity as a guide to allocate different modal combinations (e.g., unimodal, bimodal, or trimodal) to expert by attraction and repulsion routing (Sec3.4). Next, for the different after-fused results, Reliability-Aware Expert Selection (RAES) establishes emotion prototype proximity estimator to conduct comprehensive evaluation of angular-confidence and competitive intensity for reliability selection. Finally, the multimodal representation with the weights of reliability selection matrix is performed to boost MSA performance (Sec3.5).

3.2 Problem definition

Following with prior research (Zeng et al., 2021; Yu et al., 2021b) in MSA task, we aim to predict the sentiments intensity in videos by utilizing multimodal inputs. For the given utterance, the inputs data consists of text (t), vision (v) and audio (a) modalities. The sequences of three modalities are represented as, $X_v \in \mathbb{R}^{T_v \times d_v}$, $X_t \in \mathbb{R}^{T_t \times d_t}$, and $X_a \in \mathbb{R}^{T_a \times d_a}$, where $T_m, m \in t, v, a$ is the sequence length, and d_m corresponds to the embedding dimension of each modality. The objective is to predict continuous sentiment intensity values $y \in \mathbb{R}$ or a predefined set of C categories $y \in \mathbb{R}^C$.

3.3 Feature extraction and Mapping

We effectively encode the inputs with each modality $X_{m\in t,v,a}$ to obtain semantic information $F_m \in \mathbb{R}^{T_m \times d_m}$. Specifically, the BERT (Devlin et al., 2019) is used to extract text modality features F_t and the stacked transformer encoder layers[23] are leveraged to obtain the vision and audio modalities representations $F_{m\in v,a}$:

$$\begin{aligned} F_t &= \Phi_{BERT}\left(X_t\right) \in \mathbb{R}^{d_t} \\ F_m &= \Phi_{Trans}\left(X_m\right) \in \mathbb{R}^{d_m} \end{aligned} \tag{1}$$

Compared to Euclidean space, Hyperspherical embedding mitigates modality imbalance arising from feature magnitude discrepancies in similarity computation, owing to its boundedness and directional sensitivity (Ennajari et al., 2022). Normalizing features onto the unit hypersphere promotes cross-modal semantic alignment for emotional polarity modeling (Rizkallah et al., 2020; Li et al., 2025). The hypersphere space provides a scale invariant common semantic manifold for three modalities F_m, enabling framework to more reliably capture emotional polarity features. The process of mapping into hyperspherical features $h_m \in \mathbb{R}^{B\times d_s}$ by L2-normalizing is as follows:

$$h_m = \frac{F_m}{\|F_m\|_2 + \epsilon}, m \in \{t, v, a\} \tag{2}$$

where $\epsilon = 10^-8$ is a small constant for numerical stability. B and d_s are denoted as batch-size and hyperspherical space dimension, respectively

3.4 Affinity-guided Customized Modality Allocation (ACMA)

Geometry-Gradient Affinity (G²-Affinity) To resolve the first problem of combination matching, a comprehensive measurement basis should be designed to guide which modalities should be attracted together and which modalities should be repelled. Thus, we propose the G²-Affinity to commands dynamic routing of modal allocation. A high-quality modality combination should exhibit the two complementary aspects of strong gradient alignment and emotional semantic alignment. Firstly, to capture cross-modal affective consensus, we define a geometric affinity $A_{Ge}^{m,n}$ in the hyperspherical embedding space, where the cosine similarity between modalities directly reflects their emotional semantics directional coherence:

$$A_{Ge}^{m,n} = \frac{1}{2}\left(1 + h_m h_n^\top\right), \forall m, n \in \{t, v, a\}, m \neq n \tag{3}$$

Although geometric affinity $A_{Ge}^{m,n}$ effectively promotes alignment on emotional semantic level, it cannot reveal whether modalities are collaboratively optimized during training. Inspired by monitoring on gradient (Yu et al., 2020a; Wu et al., 2025b; Borsani et al., 2025), to bridge this gap, we further propose gradient affinity $A_{Gr}^{m,n}$ to capture gradient alignment for alleviating the potential gradient conflicts that may undermine learning stability and fusion efficacy:

$$\begin{aligned} \nabla_{h_m}^{\text{task}} &= \frac{\partial \mathcal{L}_{\text{task}}}{\partial h_m} \\ A_{\text{Gr}}^{m,n} &= \frac{1}{2}\left(1 + \cos\left(\nabla_{h_m}^{\text{task}}, \nabla_{h_n}^{\text{task}}\right)\right), \quad m \neq n \end{aligned} \tag{4}$$

G²-Affinity can be calculate by weighted average above two complementary component:

$$A_{m,n} = \beta A_{Ge}^{m,n} + (1-\beta) A_{Gr}^{m,n} \tag{5}$$

where $\beta \in [0, 1]$ is a hyperparameter that controls the weights of geometric affinity and gradient affinity. Higher $A_{m,n}$ represents modality combinations that maintains semantic-level alignment and optimization-level harmony.

Affinity-Guided Attraction and Repulsion Based on the guidance of G²-Affinity, the routing mechanism further acts on distributor to allocate expert with sample-customized combination. Formally, the CA-MoE consists of multiple experts, denoted as $e_1, e_2, \ldots, e_{|E|}$, where $|E|$ represents the total number of experts. In order to obtain sample-customized modality combination, we leverage an independent router $R_{e,m}, e \in \{1, 2, \ldots, |E|\}$ with corresponding learnable linear matrix $w_{e,m}$ for each modal m. Specifically, for the given embeddings h_m, the router engages the top-k-scores experts after $w_{e,m}$ and $\mathrm{softmax}$ operator. This process can be described as follows:

$$\begin{aligned} q_{e,m} &= \mathrm{softmax}(w_{e,m}(h_m)), \quad \sum_{|E|} q_{e,m} = 1 \\ R_{e,m} &= \text{Top-k}(q_{e,m}, k) \end{aligned} \tag{6}$$

For the challenge of dynamically adapting to sample-level dependency preference, the routing mechanism is required to equip with the ability to customize the allocation of modal combinations. Specifically, high-affinity modality pairs share overlapping experts and low-affinity pairs activate disjoint expert subsets. Hence, we regard the router as a situational distributor, which appropriately modulates modalities to attract or repel.

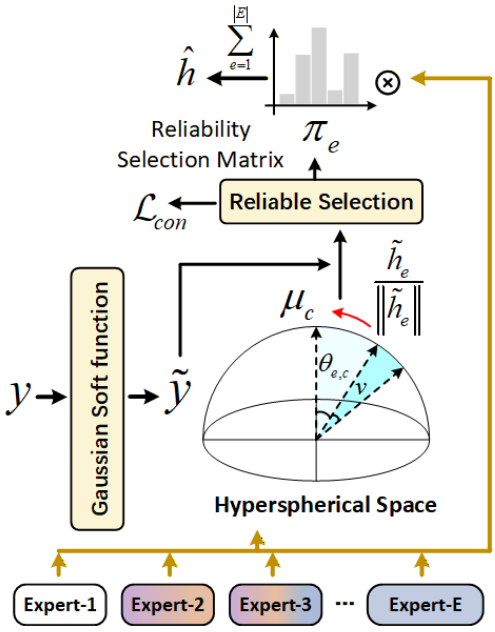


Figure 3: The details of the proposed RAES.

To achieve the allocation of similar expert subsets under attracted situation, the attraction Routing Loss $\mathcal{L}_{att}$ is designed as controller to minimizes the distance among high-affinity modal inputs q_m:

$$\mathcal{L}_{att} = \sum_{m,n} A_{m,n} \left[D_{KL}\left(q_m \,||\, q_n\right) + D_{KL}\left(q_n \,||\, q_m\right) \right] \tag{7}$$

Conversely, for the repelled situation, the Repulsion Routing Loss $\mathcal{L}_{rep}$ controls router to disperse the expert selection involving potentially conflicting modalities by maximizing divergence among low-affinity modal inputs q_m:

$$\mathcal{L}_{rep} = -\sum_{m,n} (1 - A_{m,n}) D_{JS}\left(q_m \,||\, q_n\right) \tag{8}$$

where $m, n \in \{t, v, a\}, m \neq n$. The $D_{KL}\left(\cdot \,||\, \cdot\right)$ and $D_{JS}\left(\cdot \,||\, \cdot\right)$ are Kullback-Leibler and Jensen-Shannon divergence, respectively.

Through situational distributor controlled by $\mathcal{L}_{att}$ and $\mathcal{L}_{rep}$, we achieve adaptive dependency preference allocation on sample-level by the guidance of affinity as $\tilde{h_e} \in \mathbb{R}^{B \times d_s}$:

$$\tilde{h}_e = \sum_{m} R_{e,m} h_m \tag{9}$$

3.5 Reliability-Aware Expert Selection (RAES)

Although different modal combinations have been assigned to various experts, determining the reliability of after-fused modality features and converting them into contributions to the final task remains the hanging challenge. To this end, we introduce the RAES, as shown in Fig 3, that evaluates and refine expert outputs before final prediction.

The hyperspherical space provides a geometrically appropriate embedding space for MSA task, which is conducive to angular distance polarity discrimination and enables the marginal loss (Deng et al., 2019, 2020) to sharpen class boundaries and improve discriminative robustness.

From the angular-based perspective, our establishment of emotion prototype proximity estimator begins with a learnable category center μ_c with the label $y \in \mathbb{R}^C$. In hypersphere space, expert proximity scores $p_e\left(c\right)$ are formulated as:

$$\begin{aligned} \cos\theta_{e,c} &= {\mu_c}^\top \frac{\tilde{h}_e}{\left\|\tilde{h}_e\right\|_2} \\ z_{e,c} &= \begin{cases} s \cdot \cos\left(\theta_{e,c} + \nu\right), c = y \\ s \cdot \cos\left(\theta_{e,c}\right), c \neq y \end{cases} \\ p_e\left(c\right) &= \mathrm{softmax}_c\left(z_{e,\cdot}\right) \end{aligned} \tag{10}$$

where s is the hyperparameter scaled to a numerical scale suitable for optimization. And ν is the fixed angle margin.

To adapt to the calculation of $p_e\left(c\right)$, we project discrete integer labels into a Gaussian soft target distribution with $\tilde{y} = \mathrm{softmax}\left(exp\left(-\frac{(c-y)^2}{2\sigma^2}\right)\right)$. Then, the expert angular-confidence calculated as:

$$\varpi_e = \sum^{C} p_e\left(c\right) \cdot \tilde{y} \tag{11}$$

The form of expert proximity scores $K_e = -\sum^{c} p_e\left(c\right) \log p_e\left(c\right)$ could measure the directional consensus of after-fused outputs from different experts. $H_e = \left\|\tilde{h}_e\right\|_2$ could further complementarily reflect the activation and competitive intensity among experts. Thus, reliability includes high angular agreement and magnitude responsiveness. And the final reliability selection matrix and weighted representations are defined as:

$$\begin{aligned} \pi_e &= \alpha \cdot \mathrm{softmax}\left(\gamma_1 H_e - \gamma_2 K_e\right) + \left(1 - \alpha\right) \varpi_e \\ \hat{h} &= \sum_{e=1}^{|E|} \mathrm{softmax}\left(\pi_e\right) \tilde{h}_e \end{aligned} \tag{12}$$

where γ_1 and γ_2 are hyperparameters to control the weight on the components of expert ability. The α weights the importance of expert confidence and ability in reliability selection matrix.

Table 1: Performance comparison between the CA-MoE and other baselines on the CMU-MOSI and CMU-MOSEI. The previous SOTA values are annotated with underscores, while the current SOTA values are highlighted in bold. In Acc-2 and F1-Score, the values on the left side of "/" represent "negative/non-negative", while the values on the right represent "negative/positive". A lower MAE value indicates better performance, while higher values of the other evaluation metrics (Corr, ACC-2, F1-score, ACC-7) signify superior performance.

Model	Date	CMU-MOSI					CMU-MOSEI				
		MAE↓	Corr↑	ACC-2↑	F1-score↑	ACC-7↑	MAE↓	Corr↑	ACC-2↑	F1-score↑	ACC-7↑
TFN	EMNLP,2017	0.901	0.698	-/80.8	-/80.7	34.9	0.593	0.700	-/82.5	-/82.1	50.2
MulT	ACL,2019	0.861	0.711	81.5/84.1	80.6/83.9	-	0.580	0.703	-/82.5	-/82.3	-
ICCN	AAAI,2020	0.862	0.714	-/83.1	-/83.0	39.0	0.565	0.713	-/84.2	-/84.2	51.6
MISA	ACM,2020	0.804	0.764	80.8/82.1	80.8/82.0	42.3	0.568	0.724	82.6/84.2	82.7/84.0	-
Self-MM	AAAI,2021	0.712	0.795	82.5/84.8	82.7/84.9	45.8	0.529	0.767	82.7/85.0	83.0/84.9	53.5
TETFN	PR,2023	0.717	0.800	84.1/86.1	83.8/86.1	-	0.551	0.748	84.3/85.2	84.2/85.3	-
SKEAFN	IF,2023	0.740	0.784	84.4/86.4	84.5/86.7	45.2	0.540	0.763	83.7/86.0	83.4/86.1	52.8
CRNet	KBS,2024	0.712	0.797	-/86.4	-/86.4	47.4	0.541	0.771	-/86.2	-/86.1	53.8
FGTI	ICASSP,2024	0.702	0.791	-/85.8	-/85.8	48.0	0.536	0.771	-/86.0	-/86.0	53.4
DTN	IF,2024	0.714	0.807	-/86.2	-/86.2	48.1	0.579	0.788	-/86.3	-/86.3	52.5
EMT	TAC,2023	0.705	0.798	83.3/85.0	83.2/85.0	47.4	0.527	0.774	83.4/86.0	83.7/86.0	54.5
TIEMFF	PR,2024	0.727	0.698	-/85.7	-/85.8	-	0.548	0.705	-/85.9	-/81.0	-
DLF	AAAI,2025	0.731	0.781	-/85.1	-/85.0	47.1	0.536	0.764	-/85.4	-/85.3	53.9
EMOE	CVPR,2025	0.725	0.792	83.5/85.2	83.4/85.2	45.8	0.543	0.759	82.2/85.1	82.4/84.9	53.1
MMLN	ESWA,2026	0.707	0.797	83.3/86.3	83.6/86.2	47.8	0.530	0.773	85.3/86.7	85.3/86.8	54.6
CR-GAC	ESWA,2026	0.684	0.814	85.6/87.5	85.6/87.5	44.3	0.521	0.793	84.4/86.4	84.4/86.5	54.7
CA-MoE(ours)		**0.679**	**0.822**	**86.0/88.4**	**86.2/88.5**	**48.5**	**0.515**	**0.802**	**85.8/87.5**	83.7/**87.3**	**55.4**
ΔSOTA		↓0.005	↑0.008	↑0.4/↑0.9	↑0.6/↑1.0	↑0.4	↓0.006	↑0.009	↑0.5/↑0.7	↓1.6/↑0.5	↑0.7

3.6 Output and learning Objectives

We put the weighted representations into MLP layers for linear projection:

$$\hat{y} = \text{MLP}\left(\hat{h}\right) \tag{13}$$

The model is optimized by the total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_{att} + \lambda_2 \mathcal{L}_{rep} + \lambda_3 \mathcal{L}_{con} + \lambda_4 \mathcal{L}_{bal} \tag{14}$$

where $\mathcal{L}_{task}$ represents primary task loss (Hazarika et al., 2020). Expert confidence loss $\mathcal{L}_{con} = D_{KL}\left(p_e\left(c\right) || \tilde{y}\right)$ is used to optimizes the category center. $\mathcal{L}_{bal}$ is introduced to mitigate expert under-utilization (Fedus et al., 2022). And $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ control contribution of each loss term.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate CA-MoE on three benchmark MSA datasets: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020b). The detailed statistics are summarized in Appendix A.

Implementation Details. Following standard protocols, we adopt MAE, Corr, Acc-2, F1, and Acc-7 as evaluation metrics. All experiments are conducted on an NVIDIA RTX 3060 GPU. The Appendix B shows detailed implementation settings. We also provide efficiency analysis (Appendix C) and algorithm process (Appendix D).

4.2 Comparison with State-of-the-Art

We compare CA-MoE with competitive baselines, including TFN (Zadeh et al., 2017), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), and recent SOTA methods like MMLN (Li et al., 2026) and CR-GAC (Chen et al., 2026). The quantitative results on CMU-MOSI and CMU-MOSEI are reported in Table 1, and results on CH-SIMS are shown in Table 2.

Results on CMU-MOSI and CMU-MOSEI. As observed in Table 1, CA-MoE consistently outperforms existing methods across most metrics. Notably, on the CMU-MOSI dataset, our method achieves a new state-of-the-art F1-score of 88.5% and Acc-2 of 88.4%, surpassing CR-GAC by 1.0% and 0.9%, respectively. Furthermore, CA-MoE achieves the lowest MAE (0.679) and highest correlation (0.822), demonstrating its capability to learn precise sentiment intensity. Similar improvements are observed on CMU-MOSEI, where CA-MoE maintains competitive performance, validating its robustness on large-scale data.

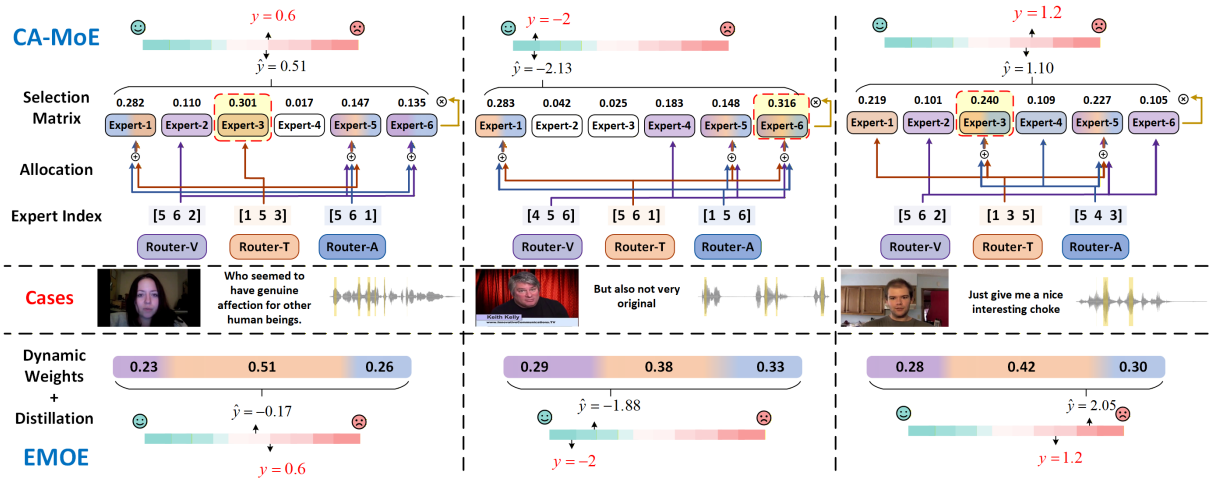


Figure 4: Qualitative comparison between CA-MoE and EMOE. The top panel visualizes the ACMA and RAES of CA-MoE, where red dashed boxes highlight high-reliability experts. The bottom panel shows EMOE's methods. CA-MoE achieves more accurate predictions by customizing expert combinations and selecting reliable presentations.

Table 2: Performance comparison on **CH-SIMS** datasets. The optimal results are highlighted in bold.

Model	MAE	Corr	Acc-2	Acc-3	Acc-5	F1
TFN	0.432	0.591	78.3	65.1	39.3	78.6
LMF	0.441	0.575	77.7	64.6	40.5	77.8
MulT	0.442	0.581	78.2	65.7	40.0	78.5
Self-MM	0.411	0.601	78.6	66.1	43.1	78.6
CENet	0.470	0.539	77.9	62.5	33.9	77.5
TETFN	0.420	0.577	81.1	63.2	41.7	80.2
EMOE	0.449	0.519	75.7	62.1	40.0	75.9
DLF	0.446	0.563	80.1	64.3	39.8	78.8
MMLN	0.406	0.602	80.7	67.1	44.6	80.9
CR-GAC	0.412	0.613	80.7	66.3	43.9	80.9
CA-MoE(ours)	**0.398**	**0.624**	**81.5**	**68.0**	**45.5**	**81.6**
ΔSOTA	↓0.008	↑0.011	↑0.8	↑0.9	↑0.9	↑0.7

Results on CH-SIMS. To verify the generalization ability on different languages, we report results on the Chinese dataset CH-SIMS in Table 2. CA-MoE achieves the best performance in MAE (0.398) and Corr (0.624), significantly outperforming the runner-up MMLN (MAE 0.406). Moreover, our method demonstrates substantial gains in multi-class classification, improving Acc-5 and Acc-3 by 0.9% compared to the previous best results.

4.3 Qualitative Analysis

To intuitively demonstrate the advantages of CA-MoE in **sample-level dependency preferences**, Figure 4 compares its performance with the baseline model EMOE on three representative samples. Although methods such as EMOE achieve flexible modal interaction through dynamic weight mechanisms and can effectively utilize all available modal information, their fusion strategies essentially still adopt a **fixed interaction paradigm**, making it difficult to precisely adapt to the modal dependency characteristics of different samples, particularly as in Cases 1 and 3. In contrast, CA-MoE allocates modality combinations under affinity guidance and further prioritizes high-scoring representations. CA-MoE identifies and prioritizes key experts (e.g., Expert-3 highlighted in the red dashed box) while avoiding falling into the fixed paradigm, thereby rectifying polarity errors and outputting predictions closer to the ground truth.

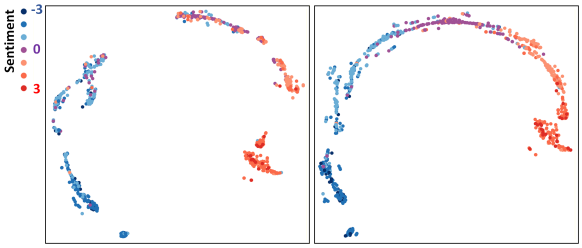
4.4 Ablation Study

Impact of Key Components. To verify the effectiveness of the model components, we conducted ablation studies on the CMU-MOSI dataset (in Table 3). The results indicate that removing any module leads to a decline in performance. Most notably, the impact of RAES as the **validator** is significant; its removal causes the MAE to rise from 0.679 to 0.760 and Acc-7 to drop by 4.8%, demonstrating the core role of reliability assessment. Furthermore, the exclusion of $\mathcal{L}_{att}$ or $\mathcal{L}_{rep}$ also results in decreased metrics, confirming the necessity of G²-Affinity in guiding modality routing.

Visualization of feature distribution. To validate the impact of the Reliability-Aware Expert Selection (RAES) module, we visualize the feature distributions of the ablation variant (w/o RAES) and the full CA-MoE in (in Fig. 5). As observed, the full CA-MoE exhibits a distinct, continuous manifold structure where samples transition

Table 3: **Ablation study of different components on CMU-MOSI dataset.** "w/o" denotes the removal of a specific module from the full model. The best results are highlighted in bold.

Model	MAE↓	Corr↑	Acc-2(%)↑	F1-score(%)↑	Acc-7(%)↑
CA-MoE	**0.679**	**0.822**	**86.0/88.4**	**86.2/88.5**	**48.8**
w/o $\mathcal{L}_{att}$	0.720	0.800	84.2/86.7	84.4/86.8	46.0
w/o $\mathcal{L}_{rep}$	0.738	0.789	83.2/84.9	83.0/85.1	43.8
w/o $\mathcal{L}_{con}$	0.713	0.803	84.8/85.7	85.0/87.7	47.1
w/o RAES	0.760	0.772	82.0/83.7	82.1/83.7	44.0



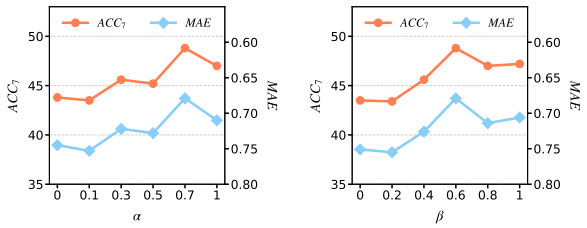
CA-MoE(*w/o RAES*) **CA-MoE**

Figure 5: t-SNE visualization of feature embeddings on the CMU-MOSI test set.

smoothly from negative (blue) to positive (red), aligning perfectly with the hyperspherical feature space design. In contrast, the distribution without RAES appears fragmented and disordered. This demonstrates that RAES effectively enhances the model's ability to learn discriminative and continuous sentiment representations.

Parameter Sensitivity Analysis. We further demonstrate that the model exhibits good robustness to hyper-parameters α and β (in Fig. 6). Experiments show that performance remains stable for α within the range of $[0.3, 0.7]$, peaking at 0.7, while β performs best around 0.6. This suggests that the model does not rely on specific parameter points but displays strong generalization stability when balancing auxiliary loss weights and the ratio of geometric to gradient affinity.

Impact of Expert Configuration. Regarding the expert settings (in Table 4), the investigation reveals that a total expert count of $|E| = 6$ serves as the optimal balance point for model capacity and generalization; too few experts lead to insufficient representation, while too many introduce redundancy. For the selection strategy, $k = 3$ proves to be the optimal solution. Compared to a single path ($k = 1$) or full activation ($k = 6$), this setting effectively leverages the complementarity of modality combinations while maintaining the advantages of sparse computation.



(a) Parameter Analysis on α (b) Parameter Analysis on β

Figure 6: **Parameter Sensitivity Analysis.** We investigate the impact of hyper-parameters α and β on the CMU-MOSI dataset.The results indicate that the model is robust within a reasonable range.

Table 4: **Ablation Study on CMU-MOSI Dataset.** We investigate the impact of the number of experts ($|E|$) and the Top-k selection strategy. The results are reported in terms of MAE, Correlation (Corr), Acc-2, F1-score, and Acc-7. The default setting is $|E| = 6$ and $k = 3$.

Setting	MAE↓	Corr↑	Acc-2(%)↑	F1-score(%)↑	Acc-7(%)↑
Impact of Expert Number ($k = 3$)					
$\|E\| = 4$	0.716	0.806	84.3/85.2	84.1/85.5	45.7
$\|E\| = 6$	**0.679**	**0.822**	**86.0/88.4**	**86.2/88.5**	**48.8**
$\|E\| = 8$	0.704	0.810	84.5/86.1	85.2/86.6	46.2
$\|E\| = 10$	0.742	0.791	83.0/85.2	82.9/84.5	44.1
$\|E\| = 12$	0.767	0.782	82.7/84.6	82.7/84.4	43.8
$\|E\| = 16$	0.762	0.776	82.2/84.0	82.5/84.1	43.7
Impact of Top-k Selection ($\|E\| = 6$)					
$k = 1$	0.721	0.811	83.8/85.7	83.7/85.2	46.2
$k = 2$	0.727	0.802	83.2/85.3	83.4/85.2	45.8
$k = 3$	**0.679**	**0.822**	**86.0/88.4**	**86.2/88.5**	**48.8**
$k = 4$	0.701	0.817	85.5/87.7	85.3/87.2	47.5
$k = 5$	0.692	0.819	85.8/87.9	86.0/88.2	48.0
$k = 6$	0.705	0.808	85.2/87.3	85.2/87.7	47.7

5 Conclusion

This paper explores the core issue of *How to dynamically identify and leverage sample-level dependency preferences among different modality combinations?* in the MSA task. Furthermore, we decompose it into two subproblems: combination matching and fusion validation. To this end, we propose a sample-level and dynamical MoE-based method, the Customized-Allocation Mixture-of-Experts (CA-MoE), which is designed for reliable modalities combination allocation. First, CA-MoE utilizes Geometry-Gradient Affinity as the attraction–repulsion routing guidance to solve customized-combination matching. Then, the validator evaluates the reliability of the fused representation for yielding a weighted final prediction. The effectiveness of CA-MoE has been thoroughly validated against MAS task datasets. We hope this work will provide inspiration for the future research on multimodal interaction paradigm.

Limitations

Although CA-MoE achieves significant performance improvements and computational efficiency, there are still some limitations that need to be addressed in future work. Firstly, due to constraints in computational resources, we did not re-implement the baseline methods but directly referenced the experimental results reported in their original papers. To ensure the fairness of the comparison, we strictly adopted the exact same dataset splits, feature extraction protocols, and evaluation metrics as those of prior studies. Secondly, while this paper effectively resolves cross-modal conflicts via the affinity-guided routing mechanism, the model currently relies on pre-extracted features (e.g., BERT). In real-world environments with high background noise or occlusion, the quality of these unimodal features may degrade, potentially affecting the calculation of geometric and gradient affinities. This imposes higher demands on feature robustness and offers opportunities for future optimization. Finally, the proposed routing mechanism implicitly learns sample-specific modality preferences through constrained modality distributions. While effective, this paradigm remains latent; future work should aim to explicitly model and interpret the underlying criteria governing modality combination decisions.

References

Garvit Ahuja, Alireza Alaei, and Umapada Pal. 2025. A new multimodal sentiment analysis for images containing textual information. *Multimedia tools and applications*, 84(21):23745–23774.

Tadas Baltruaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Yongtang Bao, Xin Zhao, Peng Zhang, Yue Qi, and Haojie Li. 2025. Hian: A hybrid interactive attention network for multimodal sarcasm detection. *Pattern Recognition*, 164:111535.

Thomas Borsani, Andrea Rosani, Giuseppe Nicosia, and Giuseppe Di Fatta. 2025. Gradient similarity surgery in multi-task deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 95–111. Springer.

Haoran Chen, Jiapeng Liu, Zuhe Li, Yushan Pan, Hongwei Tao, Huaiguang Wu, Yunyang Wang, and Chenguang Yang. 2026. Cr-gac: Cross-modal recombination via graph-attention collaborative optimization for multimodal sentiment analysis. *Expert Systems with Applications*, 298:129805.

Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. 2020. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Hafsa Ennajari, Nizar Bouguila, and Jamal Bentahar. 2022. Knowledge-enhanced spherical representation learning for text classification. In *Proceedings of the 2022 SIAM international conference on data mining (SDM)*, pages 639–647. SIAM.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. 2025. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14314–14324.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. 2024. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14755–14766.

Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. 2024. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9650–9659.

Peizhu Gong, Jin Liu, Xiliang Zhang, Xingye Li, Lai Wei, and Huihua He. 2026. Towards robust sentiment analysis with multimodal interaction graph and hybrid contrastive learning. *Pattern Recognition*, 169:111870.

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*, pages 1122–1131.

Xilin He, Haijian Liang, Boyi Peng, Weicheng Xie, Muhammad Haris Khan, Siyang Song, and Zitong Yu. 2025. Msamba: Exploring multimodal sentiment analysis with state space models. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 39, pages 1309–1317.

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*, 285:111346.

Xiang Li, Zhiqiang Dong, Xianfu Cheng, Dezhuang Miao, Haijun Zhang, Tianbo Wang, Xiaoming Zhang, and Zhoujun Li. 2026. A multi-scale representation and multi-level decision learning network for multimodal sentiment analysis. *Expert Systems with Applications*, 297:129341.

Xinyu Li, Wenqing Ye, Yueyi Zhang, and Xiaoyan Sun. 2024. Grace: Gradient-based active learning with curriculum enhancement for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5702–5711.

Zuhe Li, Panbo Liu, Yushan Pan, Weiping Ding, Jun Yu, Haoran Chen, Weihua Liu, Yiming Luo, and Hao Wang. 2025. Multimodal sentiment analysis based on disentangled representation learning and cross-modal-context association mining. *Neurocomputing*, 617:128940.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Sandra Rizkallah, Amir F Atiya, and Samir Shaheen. 2020. A polarity capturing sphere for word to vector representation. *Applied Sciences*, 10(12):4386.

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*, volume 2019, page 6558.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188.

Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. 2025a. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1601–1609.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. 2025b. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8496–8504.

Qinfu Xu, Yiwei Wei, Chunlei Wu, Leiquan Wang, Shaozu Yuan, Jie Wu, Jing Lu, and Hengyang Zhou. 2025. Towards multimodal sentiment analysis via hierarchical correlation modeling with semantic distribution constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21788–21796.

Hao Yang, Yanyan Zhao, Yang Wu, Shilong Wang, Tian Zheng, Hongbo Zhang, Zongyang Ma, Wanxiang Che, Shijin Wang, Si Wei, and 1 others. 2025a. Large language models meet text-centric multimodal sentiment analysis: A survey. *Science China Information Sciences*, 68(10):1–29.

Hao Yang, Yanyan Zhao, Yang Wu, Shilong Wang, Tian Zheng, Hongbo Zhang, Zongyang Ma, Wanxiang Che, Shijin Wang, Si Wei, and 1 others. 2025b. Large language models meet text-centric multimodal sentiment analysis: A survey. *Science China Information Sciences*, 68(10):1–29.

Yang Yang, Xunde Dong, and Yupeng Qiang. 2025c. Mse-adapter: A lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25642–25650.

Hongqi Yu, Fei Tang, Lei Zhang, Randy Gomez, Eric Nichols, and Guangliang Li. 2024. Improving perceived emotional intelligence of embodied chatbot haru via multi-modal interaction. In *2024 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 51–58. IEEE.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020a. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020b. 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*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021a. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021b. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

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*, pages 1103–1114.

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):82–88.

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)*, pages 2236–2246.

Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1262–1274.

Kai Zhao, Mingsheng Zheng, Qingguan Li, and Jianing Liu. 2025a. Multimodal sentiment analysis-a comprehensive survey from a fusion methods perspective. *IEEE Access*.

Xianbing Zhao, Xuejiao Li, Ronghuan Jiang, and Buzhou Tang. 2025b. Decoupled cross-attribute correlation network for multimodal sentiment analysis. *Information Fusion*, 117:102897.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. 2024. Glomo: Global-local modal fusion for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1800–1809.

A Datasets

We evaluate CA-MoE on three benchmark MSA datasets: CMU-MOSI, CMU-MOSEI, and CH-SIMS. The detailed statistics are summarized in Table 5. **CMU-MOSI** is a widely used dataset containing 2,199 video segments (1,284 for training, 299 for validation, and 686 for testing) annotated with sentiment intensity scores. **CMU-MOSEI** is a large-scale dataset consisting of 22,856 movie review clips, which are split into 16,326, 1,871, and 4,659 samples for training, validation, and testing, respectively. **CH-SIMS** is a Chinese multimodal sentiment analysis dataset containing 2,281 refined video segments with imbalanced modality annotations.

Table 5: Datasets details.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	299	686	2199
CMU-MOSEI	16326	1871	4659	22856
CH-SIMS	1368	456	457	2281

B Implementation Details.

Hyper-parameter	MOSI	MOSEI	SIMS
β	0.6	0.6	0.6
$\|E\|$	6	8	6
$Top-k$	3	3	3
s, ν	30,0.3	30,0.3	30,0.3
$\alpha, \gamma_1, \gamma_2$	0.7,0.8,30	0.7,0.8,30	0.7,0.8,30
$\lambda_1, \lambda_2, \lambda_3, \lambda_3$	1,1,0.1,0.01	1,1,0.1,0.01	1,1,0.1,0.01
Batch size	64	16	32
Epoch	100	100	100
Optimizer	Adam	Adam	Adam
Learning rate	1e-4	1e-4	1e-4

Table 6: Hyper-parameters setting.

Following standard protocols, we adopt MAE, Pearson Correlation (Corr), Acc-2, F1-score, and Acc-7 as evaluation metrics. All experiments are implemented on the PyTorch framework using an NVIDIA RTX 3060 GPU. The models are trained for 100 epochs using the Adam optimizer with a learning rate of $1e^{-4}$. The batch sizes are set to 64, 16, and 32 for CMU-MOSI, CMU-MOSEI, and CH-SIMS, respectively. Key hyperparameters, such as the number of experts $|E|$ and Top-k selection, vary by dataset and are detailed in Table 6.

C Efficiency Analysis

Table 7 compares the computational overhead and performance of the proposed CA-MoE against the state-of-the-art EMOE on the CMU-MOSI dataset. CA-MoE demonstrates superior efficiency, achieving a significantly faster runtime per epoch (16 s vs. 30 s) and requiring less than half the parameters (122 M vs. 317 M). Remarkably, this lightweight design facilitates better performance rather than compromising it; CA-MoE yields a lower Mean Absolute Error (0.679 vs. 0.725), proving that our model structure achieves an optimal balance between computational efficiency and predictive accuracy.

Table 7: Computational overhead.

Model	Parameters↓	Time / Epoch↓	MAE↓
EMOE	317 M	30 s	0.725
CA-MoE	**122 M**	**16 s**	**0.679**

D Algorithm Process

In this work, we present the core computational pipeline of our proposed Customized-Allocation Mixture-of-Experts (CA-MoE) using pseudocode to enhance clarity and reproducibility. Specifically, Algorithm 1 delineates the full forward inference process—from raw multimodal inputs to customized modality allocation. Concurrently, Algorithms 2 focuses on confidence estimation and loss formulation.

Together, Algorithms 1 and 2 constitute the algorithmic backbone of our approach, jointly addressing the issue of how to dynamically identify and leverage sample-level dependency preferences among different modality combinations.

Algorithm 1 Feature Extraction and Modality Allocation

Require: Text X_t, Vision X_v, Audio X_a
Ensure: $\tilde{\mathbf{h}}_e$

Feature Extraction
$\mathbf{F}_t \leftarrow \text{BERT}(X_t))$
$\mathbf{F}_v \leftarrow \text{Transformer}(X_v)$
$\mathbf{F}_a \leftarrow \text{Transformer}(X_a)$
Hyperspherical Mapping
$\mathbf{h}_m \leftarrow \text{LayerNorm}(\mathbf{F}_m), m \in \{t, v, a\}$
Routing
$\mathbf{q}_{e,m} \leftarrow \text{softmax}(w_{e,m}\mathbf{h}_m), e \in \{|E|\}$
$\mathbf{R}_{e,m} \leftarrow \text{Top-k}(q_{e,m}, k)$
Modality Allocation
for Experts $e = 1$ to $|E|$ **do**
  $\tilde{\mathbf{h}}_e \leftarrow \text{MLP}_e(\sum_m R_{e,m} \cdot \mathbf{h}_m)$
end for

Algorithm 2 Reliability-Aware Expert Selection and Loss

Require: $\tilde{\mathbf{h}}_e$, Label y
Ensure: $\hat{y}$, $\mathcal{L}_{att}$, $\mathcal{L}_{rep}$, $\mathcal{L}_{con}$
Angular-Validation
if y != None **then**
  $\tilde{y} \leftarrow \mathrm{softmax}(\text{GaussianSoftLabel}(y))$
end if
for $e = 1$ to $|E|$ **do**
  $\mu_c \leftarrow \mathcal{L}_{con}$
  $\cos\theta_{e,c} \leftarrow \mu_c^{\top} \text{LayerNorm}(\tilde{\mathbf{h}}_e)$
  if Training **then**
    $\mathbf{z}_{e,c} \leftarrow s \cdot \cos(\theta_{e,c} + \nu)$
  else
    $\mathbf{z}_{e,c} \leftarrow s \cdot \cos(\theta_{e,c})$
  end if
  $\mathbf{p}_e \leftarrow \text{softmax}(\mathbf{z}_{e,c})$
end for
$\varpi_e \leftarrow \sum^{C} p_e(c) \cdot \tilde{y}$
$K_e \leftarrow -\sum^{c} p_e(c) \log p_e(c)$
RAES
$H_e \leftarrow \|\hat{\mathbf{h}}_e\|_2$
$\pi_e \leftarrow H_e, K_e, \varpi_e$
$\hat{h} \leftarrow \sum^{|E|} \mathrm{softmax}(\pi_e)\tilde{\mathbf{h}}_e$
$\hat{y} \leftarrow \text{MLP}(\hat{h})$
Loss
$\mathcal{L}_{\text{task}} \leftarrow \text{MSE}(\hat{y}, y)$
for $m, n \in \{t, v, a\}, m \neq n$ **do**
  $A_{\text{Ge}}^{(m,n)} \leftarrow h_m h_n^{\top}$
  $A_{\text{Gr}}^{(m,n)} \leftarrow \mathcal{L}_{\text{task}}(m, n)$
  $A_{\text{m,n}} \leftarrow A_{\text{Ge}}^{(m,n)}, A_{\text{Gr}}^{(m,n)}$
  $\mathcal{L}_{\text{att}} \leftarrow A_{\text{m,n}}\text{KL}(\mathbf{q}_n, \mathbf{q}_m)$
  $\mathcal{L}_{\text{rep}} \leftarrow -(1 - A_{\text{m,n}})\text{JS}(\mathbf{q}_n, \mathbf{q}_m)$
end for
$\mathcal{L}_{\text{con}} \leftarrow \text{KL}(\mathbf{p}_e, \tilde{y})$