Group-Aware Multi-Scale Ensemble Learning for Test-Time Multi-Modal Sentiment Analysis

Kai Tang[1], **Yixuan Tang**[3*], **Tianyi Chen**[1], **Haokai Xu**[1], **Qiqi Luo**[2*],
Jin Guang Zheng[2], **Zhixin Zhang**[2], **Gang Chen**[1], **Haobo Wang**[1],
[1]Zhejiang University
[2]Ant Group
[3]National University of Singapore
kai.t,tiannychen,haokai_xu,cg,wanghaobo@zju.edu.cn
luoqiqi.lqq,zhengjinguang.zhen,zhangzhixin.zzx@antgroup.com

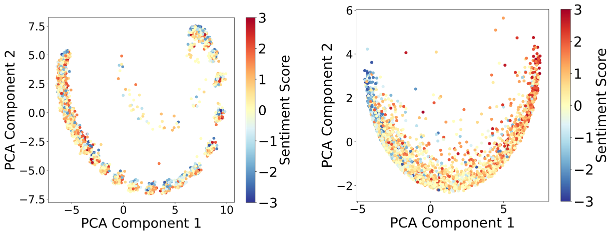
Abstract

Multi-modal Sentiment Analysis (MSA) enables machines to perceive human sentiments by integrating multiple modalities such as text, video, and audio. Despite recent progress, most existing methods assume distribution consistency between training and test data—a condition rarely met in real-world scenarios. To address domain shifts without relying on source data or target labels, Test-Time Adaptation (TTA) has emerged as a promising paradigm. However, applying TTA methods to MSA faces two challenges: a representation bottleneck inherent to the regression formulation and the inconsistency in modality fusion caused by modality-specific data augmentation techniques. To overcome these issues, we propose Group-aware Multiscale Ensemble Learning (GMEL), which leverages a von Mises-Fisher (vMF) mixture distribution to model latent sentiment groups and integrates a multi-scale re-dropout strategy for modality-agnostic feature augmentation, preserving fusion consistency. Extensive experiments on three benchmark datasets using two backbone architectures show that GMEL significantly outperforms existing baselines, demonstrating strong robustness to test-time distribution shifts in multi-modal sentiment analysis.

1 Introduction

With the rapid development of multi-modal learning (Liang, Zadeh, and Morency 2024; Ngiam et al. 2011) and dialogue system (Saha, Saha, and Bhattacharyya 2022; Firdaus et al. 2020), Multi-modal Sentiment Analysis (MSA)(Poria et al. 2016) has become the key for machines to perceive, recognize, and understand human sentiments and emotions through multiple modalities like text, video and audio. Although there has been significant progress in recent years (Das and Singh 2023; Tsai and Bai 2019), the effectiveness of these models greatly depends on the assumption of distribution consistency. However, it is hard to meet such a mild assumption in real-world scenarios.

Figure 1: Comparison of feature distribution on MOSI → MOSEI. The representation distribution learned by GMEL is smoother and more discriminative.



(a) Baseline

(b) GMEL

To maintain robustness against domain shifts, Test-Time Adaptation (TTA) has been proposed (Niu et al. 2022; Liang, He, and Tan 2025; Chen et al. 2021; Nejjar, Wang, and Fink 2023) that overcomes the distribution gaps between source and target domains during test time without access to the source data and the labels of the target data. However, most existing Test-Time Adaptation (TTA) frameworks are hard to apply to multi-modal regression problems for two reasons. First, a great number of works (Wang, Shelhamer, and Liu 2021; Wang et al. 2022; Chen et al. 2022) focus on the classification task relying on probabilistic models to generate and filter the pseudo label, which is not applicable to regression tasks like MSA. Second, majority of works (Zhang, Levine, and Finn 2022; Feng et al. 2023) are based on the assumption of uni-modality and take use of modality-specific data augmentation techniques, which fail to address the modality-fusion problem. Currently, only CASP (Guo et al. 2025a) proposed a TTA method designed for multi-modal regression by enforcing consistency and minimizing empirical risk, but it suffers from error propagation due to the two-stage framework.

Building on the above observation, the TTA for multi-modal sentiment analysis faces two critical challenges. (i)-*Representation bottleneck caused by the regression formulation.* In this setup, the dataset only indicates sentiment positivity or negativity through numerical values, while psychology-related research (Mehrabian 1996; Roy and Das 2023; Pang, Lee, and Vaithyanathan 2002) highlights that sentiment involves fine-grained cognitive distinctions measured in the Valence-Arousal-Dominance (VAD) space. This mismatch limits the source model's ability to capture latent sentiment semantics. As shown in Figure 1, when distribu-

[*]Corresponding author.

tion shifts occur, the model may misinterpret subtle sentiment cues and produce intertwined feature distribution. (ii)-*Conflict between modality fusion consistency and data augmentation.* In a multi-modal TTA task, the fusion layer is pretrained on complete multi-modal data in the source domain, and its parameters can not be altered in the adaptation process. However, most existing methods rely on the random modality elimination (e.g., CASP (Guo et al. 2025b)) to improve robustness, which destroys the completeness of the modality. Critically, this discrepancy disrupts the modality alignment learned from the source domain data, leading to performance degradation during adaptation. Thus, how to improve model robustness while maintaining modality integrity is an important issue. Both these challenges hinder the robustness and generalization of multi-modal fused representations under domain shifts.

To address these issues, we propose **G**roup-aware **M**ultiscale **E**nsemble **L**earning (dubbed **GMEL**), which ensembles multiscale features to encapsulate the intrinsic group semantics and learn discriminative representations. Specifically, it has two main components: First, inspired by psychological studies (Mehrabian 1996; Roy and Das 2023), we hypothesize that human sentiments naturally contain latent group semantics (e.g., surprise and joy tend to be positive, while disgust and sadness tend to be negative), and the relational structure among these latent concepts remains stable across domain shifts. To capture this, we model the latent sentiment groups as a von Mises-Fisher (vMF) mixture distribution, where each component relates to some one fine-grained sentiment concept. By enforcing angular concentration on the hypersphere, the vMF formulation naturally encourages semantic coherence within each group, providing a geometrically meaningful representation space. Building upon this structure, we derive self-supervised signals from grouping patterns, which serve as supervisory cues to refine feature representations during test-time adaptation. As a result, distribution-shifted samples are progressively aligned with their underlying sentiment concepts, reducing semantic ambiguity and enhancing prediction robustness under domain shifts. Second, we introduce a multi-scale re-dropout strategy that resamples dropout masks at multiple network depths to preserve modality integrity while effectively capturing collaborative latent group semantics across modalities. Finally, the synergy between the two modules yields features that capture both complete-modality interactions and latent sentiment semantics, thereby exhibiting strong robustness to test-time distribution shifts.

We comprehensively evaluate GMEL across three major MSA benchmark datasets under five cross-dataset shift scenarios using both late and early fusion backbones. Specifically, our visualized analysis in Figure1 shows that GMEL learns smoother feature distributions while maintaining connectivity between different sentiment clusters to preserve the characteristics of regression tasks [1]. Quantitatively, across 10 different experimental settings, GMEL achieves consistent improvements over the strongest baseline: average absolute gains of **2.53%** in ACC, **3.70%** in F1, and **0.05** reduction in MAE. These results collectively validate the broad effectiveness of GMEL for multi-modal sentiment regression under distribution shifts.

2 Related Work

Multi-modal Sentiment Analysis (MSA). MSA integrates information from diverse modalities, such as language, video, and audio to predict sentiment intensity (Ali and Hughes 2023; Ezzameli and Mahersia 2023; Gandhi et al. 2023). The challenges of MSA tasks lie in two aspects: representation learning and feature integration. Representation learning methods (Guo et al. 2022; Sun et al. 2023) enhance cross-modal interactions to learn modality-aligned representations. Feature-integrating strategies can be categorized into two types: feature-level fusion (early fusion) and decision-level fusion (late fusion). Feature-level fusion methods (Liang et al. 2018; Wang et al. 2019; Nagrani et al. 2021) combine different modality features to form a unified representation through direct concatenation or attention-based transformation and then feed into the projector. Decision-level fusion methods (Tsai et al. 2019; Bagher Zadeh et al. 2018) independently model modality-specific representations and subsequently integrate them into a joint decision space. Recent advances such as MMIM (Han, Chen, and Poria 2021) (hierarchical alignment), UniMSE (Hu et al. 2022) (cross-task knowledge sharing), and WisdoM (Wang et al. 2024) (contextual knowledge integration) have effectively improved performance. However, all frameworks operate under the critical assumption of domain-invariant data distributions, which become ineffective under TTA scenarios.

Test Time Adaptation (TTA). TTA aims to adapt pretrained models to domain-shifted test data without access to source samples or target labels. Significant progress has been made in uni-modal TTA: TENT (Wang, Shelhamer, and Liu 2021) adapts models by updating normalization layers via entropy minimization; SHOT (Liang, Hu, and Feng 2020) combines entropy minimization with pseudo-labeling for representation alignment; IST (Ma 2024) leverages graph learning to generate high-quality pseudo labels. Recent efforts have extended TTA to multi-modal settings: MM-TTA (Shin et al. 2022) introduces intra- and inter-modal modules for reliable pseudo labeling, while READ (Yang et al. 2024) designs robust loss objectives and attention-based fusion. However, TTA for regression remains underexplored compared to classification, which typically relies on probabilistic modeling. Existing works (Roy et al. 2023; Adachi et al. 2025) explore regression from spatial alignment but are limited to visual tasks and do not consider multi-modal settings. CASP (Guo et al. 2025b) is the first to address multi-modal regression, adopting an offline framework that processes all target samples simultaneously. However, its two-stage framework fundamentally limits adaptation quality: Stage-one feature errors propagate to pseudo-labels, and decoupled optimization misses criti-

[1] Note that latent sentiment groups do not represent strict boundaries, as regression task is inherently difficult to achieve rigid clustering. Instead, the learned features show smooth transitions between sentiment degrees.

cal synergy between feature refinement and label generation. Our work operates within the same TTA setting but learns robust feature representations in an end-to-end manner by exploiting latent sentiment semantic structures.

3 Method

3.1 Problem Statement

Taking three common modalities text, video and audio as a showcase for clarity of presentation, we formalize the MSA problem under TTA setting as follows: The datasets are divided into the source domain dataset $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_s}$ and the target domain dataset $\mathcal{T} = \{\boldsymbol{x}_i\}_{i=1}^{N_t}$ where N_s, N_t indicate the number of data in source domain and target domain respectively and $\boldsymbol{x}_i = (\boldsymbol{x}_i^t, \boldsymbol{x}_i^v, \boldsymbol{x}_i^a)$ represents multi-modal input of text, video, and audio modality. Following the offline TTA setting (Guo et al. 2025b), we assume access to the complete target dataset $\mathcal{T}$ during adaptation.

The model $\mathcal{F}$ comprises the encoder $\mathcal{M}$ and projector $\mathcal{P}$, the output of model is $\hat{y} = \mathcal{P}_{\theta_p}(\mathcal{M}_{\theta_m}(\boldsymbol{x}))$, where θ_p and θ_m are the parameters of $\mathcal{P}$ and $\mathcal{M}$. In the TTA setting, we first pre-train the model in source domain data $\mathcal{S}$, the optimization process can be formulated as $\theta_m^*, \theta_p^* = \arg\min_{\theta_m, \theta_p} \mathcal{L}(\hat{y}, y)$. Then we use the target domain dataset $\mathcal{T}$ for domain adaptation in test-time without access to the source domain $\mathcal{S}$. In this stage θ_m is frozen to minimize the self-supervised learning objection below: $\theta_p^* = \arg_{\theta_p} \min \mathcal{L}_{\text{tta}}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{T}$.

3.2 Group-Aware Learning for Latent Sentiment Semantics

In Multi-modal Sentiment Analysis (MSA), although sentiment labels are typically represented as coarse numerical values, the input data itself inherently contains fine-grained affective semantics. This suggests that pretrained models, even without explicit fine-grained annotations, may already encode latent group structures reflecting nuanced sentiment concepts through their learned representations[2]. Building on this assumption, we believe that the pretrained model's feature space roughly approximates a latent grouping of sentiments, where each group corresponds to a sentiment concept. To formalize this intuition, we adopt a hyperspherical representation framework inspired by the success of contrastive learning in multi-modal tasks (Liu et al. 2021; Yuan et al. 2021), assuming that features follow a mixture of von Mises-Fisher (vMF) distributions that can be seen as the hyperspherical counterpart of normal distributions (Du et al. 2024). Note that we do not aim to relearn new latent groups during test-time adaptation (TTA). Instead, our goal is to strengthen the partial fine-grained information implicitly carried by the coarse-grained sentiment labels. This enables test-time samples to better align with latent sentiment groups, thereby improving robustness in sentiment prediction under distribution shifts.

Latent Distribution Assumption. Inspired by the success of contrastive loss in representation learning (Chen et al. 2020a,b) that constrains representations to the unit hypersphere, we instead adopt von Mises–Fisher (vMF) mixtures, the hyperspherical counterpart of normal distributions (Du et al. 2024). Formally, assuming that there are C clusters in total, the probability destiny function for embedding $\boldsymbol{z} \in \mathbb{R}^d$ to latent cluster c is:

$$\begin{aligned} f(\boldsymbol{z}|\boldsymbol{\mu}_c, k_c) &= N_d(k_c) \exp(k_c \boldsymbol{\mu}_c^\intercal), \\ N_d(k_c) &= \frac{k_c^{d/2-1}}{2\pi^{d/2} I_{d/2-1}(k_c)}, \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_c$ is the mean direction of cluster c, k_c is the concentration parameter indicating tightness around $\boldsymbol{\mu}_c$, and $N_d(k_c)$ is the normalization factor. About the calculation of $N_d(k_c)$, $I_{d/2-1}(k_c)$ is the modified Bessel function of the first kind at order $d/2-1$, which is defined as $I_{d/2-1}(k_c) = \sum_{i=0}^{\infty} \frac{1}{i!\Gamma(d/2+i)} (\frac{k_c}{2})^{2i+d/2-1}$. The numerical evaluation of high-order Bessel functions is computationally intensive and unstable for small k_c. To address this problem, we employ the Miller recurrence algorithm (Olver 1964) to simplify calculations (Sra 2012) and details of this process are described in the Appendix. Based on the above assumption, we can get the kernel density estimation of $log(\boldsymbol{z}|c)$:

$$\rho(\boldsymbol{z}|c) = \log(\mathbb{E}_{\boldsymbol{z}_c \sim \text{vMF}(\boldsymbol{\mu}_c, k_c)}[\exp(\boldsymbol{z}^\intercal \boldsymbol{z}_c/\tau)]) = \log(\frac{N_d(k_c)}{N_d(k_c')}) \tag{2}$$

where $k_c' = \|k_c \boldsymbol{\mu}_c + \boldsymbol{z}/\tau\|$ and $\boldsymbol{z}_c$ is the set of features belong to cluster c. $\rho(\boldsymbol{z}|c)$ reflects the confidence that feature $\boldsymbol{z}$ belongs to cluster c from the distribution perspective. Therefore, we can generate pseudo labels for latent groups based on $\rho(\boldsymbol{z}|c)$. Specifically, we filter pseudo-labels by high-confidence selection. For each cluster $c \in [1, C]$, we select samples $\mathcal{D}_c$:

$$\mathcal{D}_c = \{(\boldsymbol{z}_i, \hat{\boldsymbol{y}}_i) | \arg_j \max \rho(\boldsymbol{z}_i|j) = c, \rho(\boldsymbol{z}_i|c) > \xi_c\}, \tag{3}$$

where ξ_c is the confidence threshold for selecting the top $R\%$ of samples in the subset predicted as cluster c. In practice, R is gradually increased during adaptation.

Estimation of C. Determining the number of latent groups C is a prerequisite for latent distribution estimation. To this end, we propose a simple yet effective method to estimate C using well-initialized features. Specifically, we initialize a large number of clusters C' (defaulting to 50), and perform k-means clustering on the aggregated features. We assume that a sufficiently large C' covers all meaningful latent groups, though some may be over-split[3]. Under this setting, suitable clusters tend to be denser, while too-segmented ones contain fewer samples. We thus estimate C by filtering out clusters with smaller size: $C = \sum_{i=1}^{C'} \mathbf{1}(N_i \geq t)$, where N_i is the size of i^{th} cluster, and $\mathbf{1}(\cdot)$ is an indicator function. We assign the threshold t as the mean value of N_i.

Distribution Parameter Updating. As observed in Eq. 1, the vMF distribution is governed by hyperparameters μ and

[2] We verified this view in the experiment with Table 5.

[3] We validate the robustness to C of our method in the experiment with Table**??**.

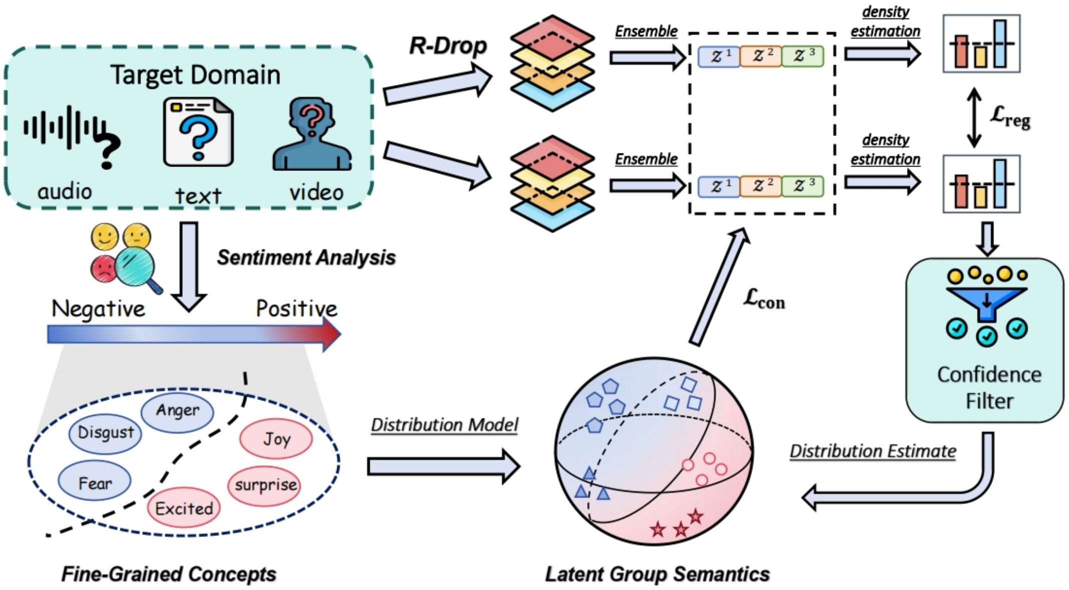


Figure 2: Overall framework of GMEL. We capture latent group semantics of sentiment to improve robustness of representations under domain shifts. To address the conflict between modality fusion consistency and data augmentation, we do multi-scale feature augmentation through r-drop and feature ensemble.

k. Leveraging the properties of the vMF distribution, we estimate the distribution parameters using only the first sample moment. This enables efficient batch-level parameter updates without requiring storage or processing of the complete feature set. Specifically, for the filtered sample set $\mathcal{D}_c$ from cluster c, we can update the sample mean $\bar{\boldsymbol{z}}_c$ through moving average manner:

$$\bar{\boldsymbol{z}}_c^t = \frac{n_c^{t-1}\bar{\boldsymbol{z}}_c^{t-1} + n_c^t \bar{\boldsymbol{z}}_c'^t}{n_c^{t-1} + n_c^t}, \tag{4}$$

where n_c^{t-1} and n_c^t respectively represent number of samples belong to class c in previous batches and current batch. $\bar{\boldsymbol{z}}_c'^t$ is the sample mean of class c in the current batch. Based on the update of $\bar{\boldsymbol{z}}_c$, we can estimate parameters $\boldsymbol{\mu}_c$ and k_c in closed form following (Sra 2012):

$$\boldsymbol{\mu}_c = \frac{\bar{\boldsymbol{z}}_c}{\|\bar{\boldsymbol{z}}_c\|_2}, k_c = \frac{\|\bar{\boldsymbol{z}}_c\|_2(d - \|\bar{\boldsymbol{z}}_c\|_2{}^2)}{1 - \|\bar{\boldsymbol{z}}_c\|_2{}^2}, \tag{5}$$

where d is the feature dimension. Having completed the estimation of the distribution parameter, we use the contrastive loss to mine latent group semantics.

Discriminative Representation Learning. Through the distribution parameter estimation described above, we can compute per-sample probability assignments to latent classes. This mitigates the scarcity of self-supervision signals in regression tasks caused by the absence of probabilistic outputs. Specifically, we integrate the probabilistic information $\hat{p}$ from the vMF mixture distribution with contrastive learning to learn discriminative feature representations.

$$\begin{aligned}\mathcal{L}_{\text{con}}^{\text{vMF}}(\boldsymbol{z}_i, \boldsymbol{y}_i) &= -\log \frac{\gamma_i \exp(\rho(\boldsymbol{z}_i|\boldsymbol{y}_i))}{\sum_{c=0}^{c=K-1} \gamma_c \exp(\rho(\boldsymbol{z}_i|\boldsymbol{y}_c)} \\ &= -\log \frac{\gamma_i(N_d(k_i)/(N_d(k_i'))}{\sum_{c=0}^{c=K-1} \gamma_c(N_d(k_c)/(N_d(k_c'))},\end{aligned} \tag{6}$$

where $\gamma_c = \frac{\sum_i \hat{p_i}^c}{N_t}$ is the prior of class c and Eq.6 is derived with Eq.2. Furthermore, we enhance contrastive learning at the feature level through kNN (Avdiukhin et al. 2024), mitigating the issue of low-quality probabilistic signals during early training stages.

$$\mathcal{L}_{\text{con}}^{\text{kNN}}(\boldsymbol{z}_i) = -\frac{1}{|\text{kNN}(\boldsymbol{z}_i)|} \log \left(\frac{\sum_{\boldsymbol{z}^+ \in \text{kNN}(\boldsymbol{z}_i)} \exp(\boldsymbol{z}_i^\intercal \boldsymbol{z}^+ / \tau)}{\sum_{\boldsymbol{z}^- \in \mathcal{T} \setminus \text{kNN}(\boldsymbol{z}_i)} \exp(\boldsymbol{z}_i^\intercal \boldsymbol{z}^- / \tau)} \right) \tag{7}$$

where τ is the temperature. Based on the above, the overall contrastive learning loss is given by,

$$\mathcal{L}_{\text{con}}(\boldsymbol{z}_i, \boldsymbol{y}_i) = \mathcal{L}_{\text{con}}^{\text{vMF}}(\boldsymbol{z}_i, \boldsymbol{y}_i) + \mathcal{L}_{\text{con}}^{\text{kNN}}(\boldsymbol{z}_i). \tag{8}$$

3.3 Multi-Scale Feature Augmentation

Many works (Tomar et al. 2023; Fleuret et al. 2021) employ data augmentation for robust adaptation. However, in multimodal TTA tasks, traditional modality-level strategies conflict with fixed fusion layers pretrained on complete modality data, disrupting the learned modality alignment patterns. Moreover, studies (Nejjar, Wang, and Fink 2023; Chen et al. 2021) show that regression transfer is particularly sensitive to feature scale variations. To this end, we introduce a novel multi-scale re-dropout augmentation strategy.

Augment Paradigm. To address the conflict between data augmentation and modality fusion consistency, we exploit the inherent stochasticity of dropout layers in neural architectures to perturb hierarchical features across different network depths through r-drop manner (Wu et al. 2021), generating multi-scale augmented representations without relying on input-level manipulations. Specifically, given input features $\boldsymbol{h}_i = \mathcal{M}_{\theta_m}(\boldsymbol{x}_i)$ and projection head $\mathcal{P}_{\theta_p}$, we

Algorithm 1: Pseudo-code of GMEL.

```
Input: Target domain dataset $\mathcal{T}$, model $\mathcal{F} = [\mathcal{M}, \mathcal{P}]$, hyperparameters $d$, $\tau$
Estimate $C$ through $\mathcal{T}$
Freeze the parameter of $\mathcal{M}$
for $epoch = 1, 2, \ldots$ do
    Divide $\mathcal{T}$ into batches
    for each bach do
        Get $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ through Multi-Scale Augmentation
        Calculate latent category probability distribution $\rho(\boldsymbol{z}|c)$ through Eq.2
        for $c = 1, \ldots, C$ do
            Get selected samples $\mathcal{D}_C$ by high confidence filtering
            Calculate $\bar{z}_c$ through Eq.4
            #Distribution Parameters Updating
            $\boldsymbol{\mu}_c \leftarrow \bar{z}_c / \|\bar{z}_c\|_2$
            $k_c \leftarrow \|\bar{z}_c\|_2 (d - \|\bar{z}_c\|_2{}^2)/(1 - \|\bar{z}_c\|_2{}^2)$
        end for
        Calculate loss $\mathcal{L}_{\text{con}}^{\text{all}}$ and $\mathcal{L}_{\text{reg}}$
    end for
    Update the parameter of $\mathcal{P}$ by minimizing $\mathcal{L}_{\text{tta}} = \mathcal{L}_{\text{con}}^{\text{all}} + \mathcal{L}_{\text{reg}}$
end for
```

forward $\boldsymbol{h}_i$ through $\mathcal{P}$ twice to extract multi-scale features from different layers of the model, resulting in two sets of multi-scale feature collections $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\mathcal{Z}'}$: $\boldsymbol{\mathcal{Z}}_i = \{\boldsymbol{z}_i^j | \boldsymbol{z}_i^j = \mathcal{P}_j(\boldsymbol{h}_i), j = 1, ..., L\}$, where $\mathcal{P}_j(\cdot)$ denotes the feature extraction at layer j of $\mathcal{P}$, and L is the total number of layers. The two passes differ in stochasticity of dropout layer, generating diverse multi-scale representations.

Feature Ensemble. Building on this strategy, we have obtained two sets of multi-scale embeddings. To improve robustness to feature-scale variations, we concatenate multi-scale features as $\boldsymbol{z}_i = [\boldsymbol{z}_i^1, \boldsymbol{z}_i^2, ..., \boldsymbol{z}_i^L], \boldsymbol{z}_i' = [\boldsymbol{z}_i^{1'}, \boldsymbol{z}_i^{2'}, ..., \boldsymbol{z}_i^{L'}]$ for computing contrastive loss:

$$\mathcal{L}_{\text{con}}^{\text{all}} = \mathcal{L}_{\text{con}}(\boldsymbol{z}_i, \hat{\boldsymbol{y}}_i) + \mathcal{L}_{\text{con}}(\boldsymbol{z}_i', \hat{\boldsymbol{y}}_i'). \tag{9}$$

On the other hand, we apply self-consistency regularization by minimizing the KL-divergence (Yu et al. 2013) between the latent class probability distributions derived from the two embeddings: $\mathcal{L}_{\text{reg}} = \sum_{k=1}^{K} \rho(\boldsymbol{z}_i|k) \log \frac{\rho(\boldsymbol{z}_i|k)}{\rho(\boldsymbol{z}_i'|k)}$. The pseudo-code of GMEL is summarized in Algorithm 1.

4 Experiments

4.1 Datasets and Evaluation Metrics

MOSI (Zadeh et al. 2016) is a multimodal sentiment dataset involving audio, text, and video modalities. It consists of 93 English YouTube videos featuring 89 distinct speakers (41 female, 48 male), where utterances are labeled on a scale of [-3 to +3]. **MOSEI** (Zadeh et al. 2018), an extension of MOSI, significantly expands both scale and diversity: it contains over 65 hours of annotated videos from 1,000+ speakers discussing 250+ topics, with 3,228 videos and 23,453 labeled segments. **SIMS** (Yu et al. 2020) provides Chinese-language multi-modal sentiment annotations across audio, text, and video modalities. It includes 60 curated videos containing 2,281 naturalistic segments labeled on a normalized [-1, +1] sentiment scale. For all three datasets, we adopt three standard metrics: (1) Binary accuracy (ACC), measuring exact prediction matches using 0 as threshold, (2) F1 score (F1), balancing precision and recall for class-imbalanced scenarios, and (3) Mean absolute error (MAE), quantifying regression performance on continuous sentiment scores.

4.2 Baselines

Regarding the baseline setting, we follow the current mainstream work (Guo et al. 2025b) and compare our methods with six baselines: **1) Source**: The source-pretrained model evaluated directly on target domain data. **2) Self-Training (ST)** : Pseudo-labels are generated by the source model and used to retrain the entire network. **3) Norm**: A parameter-efficient self-training variant (Wang, Shelhamer, and Liu 2021; Roy et al. 2023) that restricts updates to normalization layers while freezing other parameters. **4) Group Contrastive (GC)**: We adopt the GC component from TTA-IQA (Roy et al. 2023), which uses contrastive learning based on score group. The ranking loss is excluded due to its image-specific augmentation, incompatible with multi-modal inputs. **5) Reliable Fusion (RF)**: Reliable Fusion (Yang et al. 2024), which adaptively modulates cross-modal attention via confidence weights. **6) CASP**: CASP (Guo et al. 2025b) pioneers test-time adaptation for multi-modal regression, using a two-stage framework: contrastive feature alignment followed by pseudo-label refinement via consistency regularization. LLM is also a simple and effective tool in TTA scenarios. We discuss its TTA performance in specific MSA tasks in the appendix as a supplement.

4.3 Implementation Details

Raw Feature Extraction. For fair comparison, we use the same feature extraction method as the baselines. For text modality processing, we extract 768-dimensional word embeddings using pretrained BERT models: BERT-base (Devlin et al. 2019) for MOSI/MOSEI and Chinese BERT-base for SIMS. Acoustic features are extracted via LibROSA (McFee et al. 2015), while visual embeddings are obtained using OpenFace 2.0 (Baltrusaitis et al. 2018). **Backbones.** We use the encoder of transformer(Vaswani et al. 2017) as backbone and verify the effectiveness of the method with two different manners: feature-level fusion (early fusion) and decision-level fusion (late fusion). This setting is also the same as other baselines. **Training Details.** For source domain pre-training, we use the AdamW optimizer (Loshchilov, Hutter et al. 2017) with a learning rate of $1e^{-3}$. During adaptation, we employ AdamW with warm-up schedule and 0.01 weight decay. The learning rate is set to $5e^{-3}$ for experiments with late fusion and $1e^{-3}$ for experiments with early fusion. The batch size of all the experiments is 64, and the number of epochs is 15. Through estimation of C, we respectively assign $C = \{18, 19, 19\}$ to SIMS, MOSI, and MOSEI. To avoid randomness, we train the model five times using different random seeds and report the average results.

Backbone	Method	MOSEI→SIMS			MOSI→SIMS			MOSI→MOSEI			SIMS→MOSI			SIMS→MOSEI		
		ACC	**F1**	**MAE**	**ACC**	**F1**	**MAE**	**ACC**	**F1**	**MAE**	**ACC**	**F1**	**MAE**	**ACC**	**F1**	**MAE**
Early Fusion	Source	45.95	45.28	2.15	36.76	37.83	2.42	66.75	67.35	1.24	40.17	40.60	1.75	46.39	50.61	1.34
	ST	48.80	47.20	2.11	34.79	36.52	2.50	66.63	67.35	1.34	41.74	42.39	1.55	47.14	53.68	1.32
	Norm	43.76	44.06	2.25	36.23	37.94	2.47	66.98	67.54	1.30	43.95	43.40	1.56	45.80	48.13	1.29
	GC	47.64	47.60	2.10	37.22	37.70	2.29	67.12	67.40	1.22	42.67	43.08	1.54	46.77	49.12	1.30
	RF	46.12	46.25	2.18	35.18	35.97	2.46	66.84	67.39	1.27	42.58	43.02	1.59	46.61	50.39	1.31
	CASP	63.89	66.43	1.80	40.12	41.65	**2.06**	68.32	**68.90**	1.08	46.57	47.10	1.44	47.90	47.13	1.26
	GMEL	**66.49**	**74.59**	**1.74**	**44.47**	**44.61**	2.09	**68.45**	68.81	**0.96**	**47.08**	**48.64**	**1.33**	**48.52**	**48.86**	**1.26**
Late Fusion	Source	60.96	63.09	2.01	39.17	39.12	2.10	66.57	67.42	1.25	40.12	45.46	2.18	47.14	57.47	1.77
	ST	62.01	65.19	1.95	40.48	39.55	2.05	67.41	67.90	1.23	40.41	46.35	2.00	47.34	58.35	1.87
	Norm	61.40	64.38	2.04	38.51	38.84	2.12	66.62	67.53	1.30	40.27	47.22	2.30	47.70	58.74	1.84
	GC	62.62	65.38	1.98	42.23	42.89	1.97	67.03	67.83	1.25	40.94	46.87	2.21	47.45	59.07	1.76
	RF	61.12	64.07	1.97	40.19	40.01	2.06	67.11	67.70	1.28	40.18	45.98	2.28	47.61	58.44	1.86
	CASP	64.23	67.75	1.81	51.27	53.15	**1.73**	**69.12**	**69.17**	0.96	48.03	50.43	2.04	49.09	59.11	**1.60**
	GMEL	**65.97**	**69.73**	**1.77**	**56.84**	**59.03**	1.80	68.91	68.94	**0.81**	**55.45**	**62.37**	**1.81**	**51.55**	**62.85**	1.66

Table 1: Comparison across five distribution shift scenarios using two backbones. Best results are **boldfaced**, second-best performances are underlined. Values represent mean results over five random seeds.

4.4 Main Results

We evaluate methods across five domain shift settings: MOSEI→SIMS, MOSI→SIMS, MOSI→MOSEI, SIMS→MOSI, and SIMS→MOSEI. We do not use MOSEI→MOSI because MOSEI is an extension of MOSI. As shown in Table 1, GMEL achieves SOTA performance on 24 out of 30 metrics, with the remaining 6 metrics showing marginal deviations from the best results. Specifically, in terms of metric, GMEL demonstrates **2.53%** average improvement on ACC, **3.70%** improvement on F1, and **0.05** unit reduction on MAE (lower is better). We further observe that the late-fusion backbone consistently outperforms early-fusion variants across all settings, demonstrating superior suitability for MSA TTA tasks.

4.5 Ablation Study

To further analyze the contributions of different components in our method, we conduct a series of ablation studies here.

Effect of Latent Group Semantics. To investigate the effectiveness of latent group semantics, we conduct two ablation studies: 1) Removing latent group discovery (LGD): We replace the group-aware mechanism with standard NT-Xent contrastive loss (Chen et al. 2020a). 2) Alternative distribution modeling: We substitute vMF mixtures with Gaussian mixtures to assess distributional assumptions. Results in Table 2 show the effectiveness of Latent Group Semantics.

Effect of Feature Ensembling and Self-Consistency. About the effect of feature ensembling, we remove the feature ensemble mechanism and directly use final-layer embeddings z for loss computation. We also evaluate model performance without self-consistency regularization. From the results shown in Table 2, we find that 1) w/o feature ensemble shows significantly greater impact on MAE than on ACC/F1 metrics, due to MAE's sensitivity to feature-scale instability. 2)w/o $\mathcal{L}_{\text{reg}}$ exhibits more pronounced performance degradation in MOSI→SIMS experiments, likely because MOSEI's larger scale and richer semantics inherently provide stronger regularization.

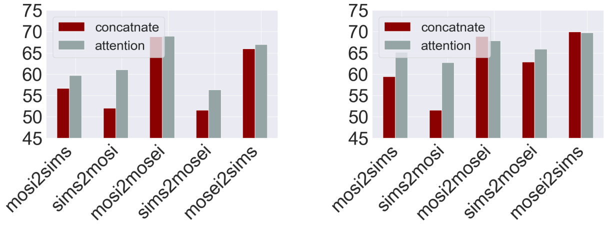
Ablation	MOSI→MOSEI			MOSEI→SIMS		
	ACC	**F1**	**MAE**	**ACC**	**F1**	**MAE**
GMEL	68.91	68.94	0.81	65.97	69.73	1.77
Complete Module Ablation						
w/o LGD	65.27	65.60	0.98	62.23	63.57	1.92
w/o Feature Ensembling	67.29	67.35	1.02	65.27	66.82	1.98
w/o $\mathcal{L}_{\text{reg}}$	65.19	66.02	0.92	64.87	68.12	1.83
w/o Data Augment	67.15	67.34	1.08	61.72	64.09	1.90
w/o $\mathcal{L}_{\text{con}}^{\text{kNN}}$	68.55	68.79	0.80	65.02	69.66	1.83
Alternative Method Comparison						
Gaussian Mixtures	67.95	67.72	0.84	64.59	67.92	1.82
Random Modality Masking	68.02	68.41	0.84	64.35	68.90	1.82

Table 2: Ablation results on MOSI→MOSEI and MOSEI→SIMS.

Effect of Augment Methods. We study the impact of augmentation strategies: (1) Data augmentation—removing it during training and also evaluating random modality masking as an alternative; (2) Contrastive learning augmentation—excluding $\mathcal{L}_{\text{con}}^{\text{kNN}}$ to assess its contribution. The ablation study shown in Table 2 validates the effectiveness of both data augmentation and $\mathcal{L}_{\text{con}}^{\text{kNN}}$, while demonstrating that preserving modality integrity through our re-dropout strategy—rather than random masking—significantly enhances robustness to distribution shifts in MSA. This confirms that maintaining complete modality interactions and multi-scale feature perception is critical for effective TTA.

4.6 Further Analysis

Sensitivity Analysis. The number of latent categories is an important hyperparameter in GMEL. We extensively



(a) Accuracy (b) F1-Score

Figure 3: Comparison between different modality fusion methods.

tested the performance changes of GMEL in different domain shifting under different C value settings, taking the MAE indicator as an example, demonstrating that GMEL is robust to the number of latent categories. As shown in the results in the table 3, when C is set in the range of 5-30, GMEL maintains SOTA or slightly lower than SOTA performance on five datasets. Additionally, we can discover that with $C = 20$, which is most similar to our estimation $\{18, 19, 19\}$, GMEL achieves relatively the best results, demonstrating the effectiveness of our estimation of C.

C	$\mathcal{D}_1$to$\mathcal{D}_2$	$\mathcal{D}_1$to$\mathcal{D}_3$	$\mathcal{D}_2$to$\mathcal{D}_1$	$\mathcal{D}_2$to$\mathcal{D}_3$	$\mathcal{D}_3$to$\mathcal{D}_2$
5	1.92	0.83	1.90	1.72	1.82
10	1.83	0.84	1.84	1.70	1.85
15	1.75	0.82	1.79	1.66	1.82
20	1.78	0.77	1.79	1.63	1.76
25	1.81	0.81	1.80	1.72	1.81
30	1.89	0.79	1.82	1.75	1.79

Table 3: Sensitivity analysis on latent category number C about MAE. For brevity, we denote the **MOSI**, **SIMS**, and **MOSEI** datasets as $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$, respectively.

Extended to pre-train. As an extension, we investigate the impact of applying Latent Group Learning to source-domain model training by evaluating: (1) source-domain performance, (2) zero-shot target-domain performance, and (3) adapted target-domain performance. As shown in Table 5, mining latent sentiment semantics yields marginal improvements in supervised source-domain training, but demonstrates stronger generalization performance on unseen target domains. This suggests that the model has already captured fine-grained sentiment concepts through source-domain training, and enhancing its sensitivity to the partial fine-grained information embedded in sentiment scores facilitates learning domain-shift-robust features.

Method	GMEL	RF	ST	Norm
MOSI→SIMS	8.42s	9.42s	8.25s	8.32s
SIMS→MOSI	5.92s	6.27s	5.81s	5.90s

Table 4: Comparison of epoch time in different settings.

Training Complexity. We report the computational time consumption of GMEL across training epochs and compare it with other methods. Notably, we exclude CASP from this comparison due to its two-stage framework's inherent computational overhead. All timing results are benchmarked on a single RTX A5000 (24GB) with batchsize 64 to ensure fair evaluation. As the results shown in Table 4, GMEL achieves lower time complexity than attention-based methods and maintain comparable computational efficiency against conventional fine-tuning techniques.

Test	Normal Pretrain			Group-Aware Pretrain		
	ACC	F1	MAE	ACC	F1	MAE
Source Domain	79.91	79.88	1.00	80.02	79.75	1.00
Target Domain	42.27	42.05	2.21	49.89	48.12	2.04
Domain Adaptation	56.67	59.26	1.82	56.84	59.03	1.80

Table 5: Performance comparison of normal and group-aware pretrain method conducted under MOSI→SIMS.

Analysis of Modality Fusion. While our main results use standard early/late fusion backbones (feature concatenation), we further validate our method's generality using attention-based fusion (Yang et al. 2024). As shown in Fig. 3, attention improves overall performance in both ACC and F1 metrics, meaning greater performance improvement prospects. Crucially, our approach maintains superior performance across all architectures, confirming its architectural robustness. For fair comparison with baselines, we report main results without attention layers, aligning with standard fusion practices.

Supplementary experiments in Appendix. To further validate the robustness and parameter sensitivity of GMEL, we conduct comprehensive ablation studies in the appendix. These analyses systematically evaluate the impact of critical hyperparameters—including dropout rate, filter ratio $\mathcal{R}$, and batch size. Additionally, we explore the potential of LLMs for MSA TTA through in-context learning experiments, revealing both strengths and limitations of this emerging approach compared to our method.

5 Conclusion

In this paper, we propose GMEL, a novel framework for robust multi-modal sentiment analysis under test-time domain adaptation. Inspired by related psychological studies, we capture fine-grained information carried by sentiment through vMF mixture distribution. Additionally, to address the conflict between data augmentation and modality fusion consistency, we design a new augment paradigm leveraging the inherent randomness of dropout layers in the model and ensembling multi-scale features. Extensive experiments on three MSA benchmarks with two backbone architectures demonstrate that GMEL outperforms baselines by significant margins across multiple metrics. We hope our work can inspire future research to further improve the test-time adaptation methods for multi-modal regression tasks.

Acknowledgements

This paper is supported by the National Regional Innovationand Development Joint Fund (No. U24A20254) and MYbank, Ant Group. Haobo Wang is also supported by the Fundamental Research Funds for the Central Universities (No. 226-2025-00085).

References

Adachi, K.; Yamaguchi, S.; Kumagai, A.; and Hamagami, T. 2025. Test-time Adaptation for Regression by Subspace Alignment. In *ICLR*.

Ali, K.; and Hughes, C. E. 2023. A unified transformer-based network for multimodal emotion recognition. *arXiv preprint arXiv:2308.14160*.

Avdiukhin, D.; Chatziafratis, V.; Fischer, O.; and Yaroslavtsev, G. 2024. Embedding Dimension of Contrastive Learning and k-Nearest Neighbors. *Advances in Neural Information Processing Systems*.

Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *ACL*.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*.

Chen, D.; Wang, D.; Darrell, T.; and Ebrahimi, S. 2022. Contrastive test-time adaptation. In *CVPR*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*.

Chen, X.; Wang, S.; Wang, J.; and Long, M. 2021. Representation Subspace Distance for Domain Adaptation Regression. In *ICML*.

Das, R.; and Singh, T. D. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Du, C.; Wang, Y.; Song, S.; and Huang, G. 2024. Probabilistic Contrastive Learning for Long-Tailed Visual Recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*.

Ezzameli, K.; and Mahersia, H. 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*.

Feng, C.-M.; Yu, K.; Liu, Y.; Khan, S.; and Zuo, W. 2023. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Firdaus, M.; Chauhan, H.; Ekbal, A.; and Bhattacharyya, P. 2020. EmoSen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing*.

Fleuret, F.; et al. 2021. Test time adaptation through perturbation robustness. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; and Hussain, A. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*.

Guo, J.; Tang, J.; Dai, W.; Ding, Y.; and Kong, W. 2022. Dynamically adjust word representations using unaligned multimodal information. In *ACMMM*.

Guo, Z.; Jin, T.; Xu, W.; Lin, W.; and Wu, Y. 2025a. Bridging the gap for test-time multimodal sentiment analysis. In *AAAI*.

Guo, Z.; Jin, T.; Xu, W.; Lin, W.; and Wu, Y. 2025b. Bridging the gap for test-time multimodal sentiment analysis. In *AAAI*.

Han, W.; Chen, H.; and Poria, S. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *EMNLP*.

Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; and Li, Y. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In *EMNLP*.

Liang, J.; He, R.; and Tan, T. 2025. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*.

Liang, J.; Hu, D.; and Feng, J. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *ICML*.

Liang, P. P.; Liu, Z.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In *EMNLP*.

Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*.

Liu, Y.; Fan, Q.; Zhang, S.; Dong, H.; Funkhouser, T.; and Yi, L. 2021. Contrastive multimodal fusion with tupleinfonce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Loshchilov, I.; Hutter, F.; et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

Ma, J. 2024. Improved self-training for test-time adaptation. In *CVPR*.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P. W.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and Music Signal Analysis in Python. In *SciPy*.

Mehrabian, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*.

Nejjar, I.; Wang, Q.; and Fink, O. 2023. Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices. In *CVPR*.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Y.; et al. 2011. Multimodal deep learning. In *ICML*.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*.

Olver, F. W. 1964. Error analysis of Miller's recurrence algorithm. *Mathematics of Computation*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Poria, S.; Cambria, E.; Howard, N.; Huang, G.-B.; and Hussain, A. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*.

Roy, B.; and Das, S. 2023. Perceptible sentiment analysis of students' WhatsApp group chats in valence, arousal, and dominance space. *Soc. Netw. Anal. Min.*

Roy, S.; Mitra, S.; Biswas, S.; and Soundararajan, R. 2023. Test time adaptation for blind image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Saha, T.; Saha, S.; and Bhattacharyya, P. 2022. Towards sentiment-aware multi-modal dialogue policy learning. *Cognitive Computation*.

Shin, I.; Tsai, Y.-H.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K.-J. 2022. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *CVPR*.

Sra, S. 2012. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of I s (x). *Computational Statistics*.

Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Tomar, D.; Vray, G.; Bozorgtabar, B.; and Thiran, J.-P. 2023. Tesla: Test-time self-learning with automatic adversarial augmentation. In *CVPR*.

Tsai, Y.; and Bai. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *ACL*.

Tsai, Y. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*.

Wang, D.; Shelhamer, E.; and Liu, S. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *CVPR*.

Wang, W.; Ding, L.; Shen, L.; Luo, Y.; Hu, H.; and Tao, D. 2024. WisdoM: Improving Multimodal Sentiment Analysis by Fusing Contextual World Knowledge. In *ACM Multimedia*.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: dynamically adjusting word representations using nonverbal behaviors. In *AAAI*.

Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; Liu, T.-Y.; et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in neural information processing systems*.

Yang, M.; Li, Y.; Zhang, C.; Hu, P.; and Peng, X. 2024. Test-time adaptation against multi-modal reliability bias. In *ICLR*.

Yu, D.; Yao, K.; Su, H.; Li, G.; and Seide, F. 2013. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *ACL*.

Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; and Faieta, B. 2021. Multimodal contrastive training for visual representation learning. In *CVPR*.

Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *ACL*.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intell. Syst.*

Zhang, M.; Levine, S.; and Finn, C. 2022. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*.