applied system innovation

MDPI

Article

MLD-FER: Multimodal Label Distribution-Guided Fusion and Expert Routing for Multimodal Sentiment Analysis

Lei Yang * and Hiroyuki Shinnou

Graduate School of Science and Engineering, Ibaraki University, 4-12-1 Nakanarusawa-cho, Hitachi 316-8511, Japan; hiroyuki.shinnou.0828@vc.ibaraki.ac.jp
* Correspondence: lainey1105yang@gmail.com

Abstract

Multimodal sentiment analysis aims to integrate textual, acoustic, and visual information for sentiment prediction. However, existing approaches commonly derive modality weighting or routing cues from learned semantic representations or other task-dependent signals, while explicit modality-level statistical information remains insufficiently explored. Moreover, existing Multimodal Label Distribution Learning mainly applies the learned distribution to direct modality weighting and fusion, leaving its potential to coordinate multiple adaptive components underexplored. To address these issues, we propose Multimodal Label Distribution-Guided Fusion and Expert Routing (MLD-FER). MLD-FER first learns a normalized, sample-dependent Multimodal Label Distribution (MLD) from modality-level statistical descriptors. The learned MLD serves as a task-adaptive, sample-dependent modality weighting signal that modulates multimodal fusion and provides additional conditioning information for sparse expert routing. In this way, the same MLD coordinates cross-modal fusion and sparse expert routing as a shared sample-dependent signal. Experiments on two widely used benchmarks, CMU-MOSI and CMU-MOSEI, show that MLD-FER achieves competitive performance across multiple evaluation metrics compared with representative existing methods. Additional analyses further characterize the behavior and contributions of the proposed mechanisms. These results support the use of modality-level statistical descriptors for constructing task-adaptive modality weighting signals in multimodal sentiment analysis.

Keywords: multimodal sentiment analysis; multimodal label distribution; cross-modal fusion; mixture-of-experts

Check for updates

Academic Editor: Sheng-Tzong Cheng

Received: 9 October 2026
Revised: 19 November 2026
Accepted: 23 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Sentiment analysis is an important research topic for understanding human opinions, attitudes, and affective expressions. Traditional sentiment analysis relies mainly on textual information. However, human sentiment is often conveyed through multiple signals, including linguistic content, vocal characteristics, and facial behaviors [1]. Multimodal sentiment analysis (MSA) therefore combines textual, acoustic, and visual information to obtain a more complete understanding of sentiment [2]. By exploiting complementary information from different modalities, MSA can alleviate the limitations of relying on a single information source and improve sentiment prediction in complex scenarios.

Existing MSA methods have extensively investigated cross-modal interaction and adaptive multimodal fusion. Recent surveys have also highlighted the rapid development of multimodal fusion strategies and text-centric multimodal modeling in MSA [1–3].

For example, MulT [4] employs directional cross-modal attention to model interactions between unaligned modality sequences, while MAG-BERT [5], MISA [6], Self-MM [7], and TETFN [8] explore multimodal representation learning and fusion from different perspectives. TGMoE [9] integrates text-guided cross-modal learning with a sparsely gated mixture-of-experts framework to enhance multimodal representations. Another related line of research considers sample-dependent modality weighting for adaptive multimodal processing. EMOE [10] learns sample-specific modality weights from multimodal features for adaptive fusion. Despite these advances, modality weighting and routing decisions are generally derived from learned semantic representations or other task-dependent latent signals, while explicit modality-level statistical descriptors remain comparatively underexplored as side-information for learning sample-dependent modality weights.

Multimodal Label Distribution Learning (MLDL) provides a related perspective by introducing side-information to learn a normalized Multimodal Label Distribution (MLD) [11]. In MLD, each description degree is associated with one modality, and the resulting distribution provides a normalized modality weighting pattern conditioned on the given side-information. Since explicit MLD annotations are unavailable, the description degrees are learned indirectly through the downstream prediction objective. Accordingly, in this work, they are treated as task-adaptive, sample-dependent modality weights rather than explicitly supervised measures of modality reliability or ground-truth importance. The original MLDL framework mainly uses the learned MLD for direct modality weighting and fusion, while its potential as a shared sample-dependent conditioning signal for coordinating multimodal fusion and sparse expert routing remains insufficiently explored.

To address these limitations, we propose Multimodal Label Distribution-Guided Fusion and Expert Routing (MLD-FER) for multimodal sentiment analysis. The key idea of MLD-FER is to learn a normalized, sample-dependent MLD from modality-level statistical descriptors and use the same MLD as a common conditioning signal for cross-modal fusion and sparse expert routing. Specifically, we first construct (1) a Multimodal Label Distribution Learning module. The module derives modality-level statistical descriptors from textual, acoustic, and visual features and uses them as side information to estimate the MLD. We then design (2) an MLD-Modulated Cross-Modal Fusion module, which combines text-guided cross-modal interaction with MLD-modulated unimodal representations to construct the fused multimodal representation. Furthermore, we introduce (3) an MLD-Informed Sparse Expert Routing module, which separately transforms the fused sample representation and the learned MLD into expert-level routing logits and combines the resulting logit terms to condition sparse Top-k expert selection. In this way, MLD-FER uses the same sample-dependent MLD to coordinate cross-modal fusion and sparse expert routing. In summary, our main contributions are as follows:

- We propose Multimodal Label Distribution-Guided Fusion and Expert Routing (MLD-FER), a unified framework that learns a sample-dependent MLD from modality-level statistical descriptors and uses the same MLD to coordinate cross-modal fusion and sparse expert routing.
- We design an MLD-Modulated Cross-Modal Fusion module and an MLD-Informed Sparse Expert Routing module. The former combines text-guided cross-modal interaction with MLD-modulated unimodal representations, while the latter incorporates both the fused sample representation and the learned MLD directly into expert-level routing logits for sparse Top-k expert selection.
- Experiments on the CMU-MOSI and CMU-MOSEI datasets show that MLD-FER achieves competitive performance. Controlled component ablations, statistical descriptor analyses, and visualization experiments further examine the effects and internal behavior of the proposed MLD-based fusion and routing mechanisms.

2. Related Work

2.1. Multimodal Sentiment Analysis

Multimodal sentiment analysis integrates textual, acoustic, and visual signals to obtain a more comprehensive representation of human sentiment. It has been widely applied in areas such as human–computer interaction [12], education [13], and healthcare [14]. By exploiting complementary information across modalities, MSA can alleviate the limitations of relying on a single information source. Nevertheless, effectively modeling cross-modal relationships and constructing informative multimodal representations remain central challenges [15,16].

Recent studies have addressed these challenges from different perspectives. MISA [6] decomposes multimodal representations into modality-invariant and modality-specific subspaces to preserve both shared and distinctive information across modalities. Self-MM [7] introduces unimodal self-supervised tasks to enhance modality-specific representation learning and provide additional supervision for multimodal prediction. MMIM [17] employs hierarchical mutual information maximization to retain task-relevant information during multimodal fusion. More recent studies have further explored text-enhanced fusion, adaptive modality weighting, and expert-based multimodal processing. TETFN [8] strengthens textual information during multimodal fusion, while TGMoE [9] combines text-guided cross-modal interaction with a sparsely gated mixture-of-experts framework, enabling experts to selectively process fused multimodal representations. EMOE [10] learns sample-specific modality weights from multimodal features and employs modality experts for adaptive fusion. MISA-GMC [18] further combines gated multimodal fusion with modality-relationship regularization to improve multimodal representation learning. These studies demonstrate the effectiveness of semantic representation learning, cross-modal interaction, and adaptive fusion for MSA. Our work complements this line of research by investigating modality-level statistical descriptors as side-information for learning task-adaptive, sample-dependent modality weights and using the resulting MLD to modulate both multimodal fusion and sparse expert routing.

2.2. Multimodal Label Distribution Learning

Label Distribution Learning (LDL) [19] extends conventional single-label learning by assigning a description degree to each candidate label, thereby representing the extent to which each label describes an instance. Building on LDL, Multimodal Label Distribution Learning (MLDL) [11] reinterprets the distribution support from class labels to modalities and introduces a Multimodal Label Distribution (MLD) in which each modality is associated with a normalized description degree. Given side-information $\mathbf{s}$, the MLD is expressed as $\mathbf{d} = [d_1, \ldots, d_M]$, where d_m denotes the description degree assigned to the m-th modality under $\mathbf{s}$, $d_m \geq 0$, and $\sum_{m=1}^{M} d_m = 1$. The resulting MLD therefore provides a side-information-conditioned distribution over modalities. Since explicit MLD annotations are generally unavailable, the distribution is learned indirectly through supervision from the downstream task.

The original MLDL framework derives side-information from modality features and uses the resulting MLD primarily for adaptive modality weighting and fusion. More recently, distribution-learning formulations have also been explored in multimodal sentiment analysis [20], with a primary focus on modeling sentiment-related distributions. Such formulations are conceptually different from MLD, whose distribution support corresponds to modalities and can therefore be used as a normalized modality weighting signal. Building on the MLDL formulation, our work treats the learned MLD as a shared sample-dependent conditioning signal and examines its use in both cross-modal fusion and sparse expert routing.

3. Method

3.1. Overview

As illustrated in Figure 1, the proposed MLD-FER framework consists of three core modules. (a) Multimodal Label Distribution Learning Module, which derives modality-level statistical descriptors from the pre-extracted numerical textual, acoustic, and visual features and uses them as side information to estimate a normalized Multimodal Label Distribution (MLD). The resulting MLD provides sample-dependent modality weighting for multimodal fusion and expert routing. (b) MLD-Modulated Cross-Modal Fusion Module, which combines text-guided cross-modal interaction with MLD-modulated unimodal representations. (c) MLD-Informed Sparse Expert Routing Module, which independently maps the fused sample representation and the learned MLD to expert-level routing logits and combines the two logit terms before sparse Top-k expert selection. The selected expert outputs are then weighted and aggregated for sentiment prediction.

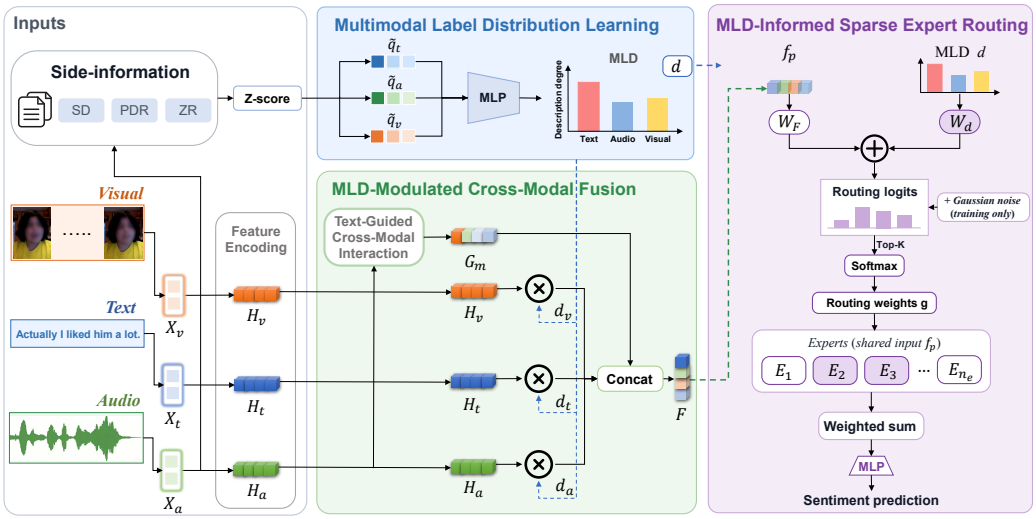


Figure 1. Overview of the proposed MLD-FER framework, which consists of three main modules: Multimodal Label Distribution Learning, which derives modality-level statistical descriptors and uses them as side-information to estimate a normalized, sample-dependent MLD; MLD-Modulated Cross-Modal Fusion, which combines text-guided cross-modal interaction with MLD-modulated unimodal representations to construct the fused multimodal representation; and MLD-Informed Sparse Expert Routing, which combines routing logits derived from the fused multimodal representation and the MLD to perform sparse Top-k expert selection for sentiment prediction.

3.2. Multimodal Label Distribution Learning Module

To capture sample-dependent modality variation from a statistical perspective, we construct a Multimodal Label Distribution Learning module. The module derives modality-level statistical descriptors from textual, acoustic, and visual features and uses them as side information to estimate a normalized Multimodal Label Distribution (MLD).

Let $\mathbf{X}_m \in \mathbb{R}^{T_m \times D_m}$ denote the pre-extracted numerical feature sequence of modality $m \in \{t, a, v\}$, where T_m and D_m denote its sequence length and feature dimension, respectively. When padded temporal positions are present, they are excluded only from the computation of the statistical descriptors, while the original feature sequences remain unchanged for subsequent processing. Let $\mathcal{T}_m$ denote the set of valid temporal positions for modality m. Following the statistical side-information formulation in [11], we use standard deviation

(SD), peak-to-dispersion ratio (PDR), and zero-rate (ZR) to characterize different statistical aspects of each modality. The statistical descriptor vector of modality m is defined as

$$\mathbf{q}_m = [\mathrm{SD}(\mathbf{X}_m),\ \mathrm{PDR}(\mathbf{X}_m),\ \mathrm{ZR}(\mathbf{X}_m)]. \tag{1}$$

For clarity, after excluding padded positions, the remaining scalar feature elements are collected as $\{x_i^m\}_{i=1}^{N_m}$, where $N_m = |\widetilde{J}_m| \cdot D_m$ denotes the total number of valid scalar feature elements and $\bar{x}_m$ denotes their mean. The standard deviation is calculated as

$$\mathrm{SD}(\mathbf{X}_m) = \sqrt{\frac{1}{N_m}\sum_{i=1}^{N_m}\left(x_i^m - \bar{x}_m\right)^2}. \tag{2}$$

Because the reference term used in the original formulation is the feature mean rather than a noise-free signal, we refer to the corresponding statistic as the peak-to-dispersion ratio (PDR) to avoid a signal-to-noise interpretation. It is defined as

$$\mathrm{PDR}(\mathbf{X}_m) = 10\log_{10}\left(\frac{M_m^2 + \epsilon}{\mathrm{SD}(\mathbf{X}_m)^2 + \epsilon}\right), \tag{3}$$

where $M_m = \max_{1\le i\le N_m} |x_i^m|$ denotes the maximum absolute magnitude among the valid feature elements, and $\epsilon = 10^{-8}$ is introduced for numerical stability. PDR therefore characterizes the peak feature magnitude relative to the dispersion of the valid feature values. The zero-rate is defined as

$$\mathrm{ZR}(\mathbf{X}_m) = \frac{1}{N_m}\sum_{i=1}^{N_m}\mathbb{I}(x_i^m = 0), \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Since padded positions are excluded before descriptor computation, ZR measures the proportion of exact zero-valued elements within the valid feature region rather than sequence padding. Together, SD, PDR, and ZR characterize feature dispersion, peak magnitude relative to dispersion, and the proportion of zero-valued feature elements, respectively. These descriptors jointly constitute the statistical side-information used for MLD estimation.

Because the numerical ranges of SD, PDR, and ZR differ and the raw feature scales are modality-dependent, each modality–descriptor pair is standardized independently across the training samples before being passed to the shared scoring network. Specifically,

$$\tilde{q}_{m,j} = \frac{q_{m,j} - \mu_{m,j}}{\sigma_{m,j} + \epsilon}, \tag{5}$$

where $\mu_{m,j}$ and $\sigma_{m,j}$ are the mean and standard deviation of the j-th descriptor for modality m, estimated across the training samples only. The resulting statistics are fixed and subsequently applied to the validation and test sets. No additional cross-modal normalization is performed. We denote the normalized descriptor vector by $\tilde{\mathbf{q}}_m$.

To map the modality-level statistical descriptors to comparable modality scores, we employ a scoring network with parameters shared across the three modalities:

$$e_m = \mathbf{w}_d^\top \phi\left(\mathbf{W}_q \tilde{\mathbf{q}}_m + \mathbf{b}_q\right) + b_d, \qquad m \in \{t, a, v\}, \tag{6}$$

where $\mathbf{W}_q$, $\mathbf{b}_q$, $\mathbf{w}_d$, and b_d are learnable parameters shared across the three modalities, and $\phi(\cdot)$ denotes a nonlinear activation function. Parameter sharing allows the normalized statistical descriptors of different modalities to be evaluated under the same scoring function. The resulting modality scores are jointly normalized to obtain the MLD:

$$\mathbf{d} = \mathrm{Softmax}([e_t, e_a, e_v]) = [d_t, d_a, d_v], \tag{7}$$

where d_m denotes the description degree assigned to modality $m \in \{t, a, v\}$, with $d_m \geq 0$ and $\sum_{m\in\{t,a,v\}} d_m = 1$. In our framework, these description degrees are treated as task-adaptive, sample-dependent modality weights rather than explicitly supervised measures of modality reliability or ground-truth importance. Together, the three description degrees form the sample-dependent MLD. Since explicit MLD annotations are generally unavailable in existing multimodal sentiment analysis datasets, the MLD is jointly learned with the entire framework through the final sentiment prediction objective. The resulting MLD is subsequently used to modulate cross-modal fusion and inform sparse expert routing.

3.3. MLD-Modulated Cross-Modal Fusion Module

Although text-guided cross-modal interaction effectively incorporates acoustic and visual cues into the textual semantic stream, the resulting fused representation does not explicitly account for sample-dependent variations in the statistical characteristics of the individual modalities. To address this limitation, the proposed MLD-Modulated Cross-Modal Fusion module combines text-guided cross-modal interaction with MLD-based modulation of the original modality-specific representations, allowing sample-dependent statistical side-information to participate in the final multimodal fusion.

Given the encoded textual, acoustic, and visual feature sequences $\mathbf{H}_t$, $\mathbf{H}_a$, and $\mathbf{H}_v$, the text-guided cross-modal interaction scheme follows TGMoE [9], which builds on the directional cross-modal Transformer of MulT [4]. The textual representation serves as the query, while the acoustic and visual representations separately provide the keys and values. The visual-to-text and audio-to-text cross-modal Transformers share the same architecture, each comprising L stacked cross-modal attention layers.

Before cross-modal interaction, the three modality representations are mapped to a common hidden dimension D. The textual representation is retained or linearly projected when necessary, while the acoustic and visual representations are projected to the same dimension using modality-specific 1×1 temporal convolutions. Each cross-modal Transformer layer employs residual connections around the attention and feed-forward transformations together with layer normalization, following the adopted backbone architecture.

Taking the visual modality as an example, the textual representation provides the queries, while the visual representation provides the keys and values. The core attention operation within the visual-to-text cross-modal Transformer is expressed as

$$\mathbf{C}_v = \mathrm{Softmax}\left(\frac{\left(\mathbf{H}_t\mathbf{W}_Q^t\right)\left(\mathbf{H}_v\mathbf{W}_K^v\right)^\top}{\sqrt{d_k}}\right)\left(\mathbf{H}_v\mathbf{W}_V^v\right), \tag{8}$$

where $\mathbf{W}_Q^t$, $\mathbf{W}_K^v$, and $\mathbf{W}_V^v$ are learnable projection matrices, and d_k denotes the dimension of the projected key vectors. Here, $\mathbf{C}_v$ denotes the cross-modal attention message extracted from the visual source according to the textual queries. The corresponding visual-conditioned textual representation is obtained through the residual and feed-forward transformations as

$$\begin{aligned} \mathbf{U}_v &= \mathbf{H}_t + \mathbf{C}_v, \\ \mathbf{A}_v &= \mathrm{LN}(\mathbf{U}_v + \mathrm{FFN}(\mathrm{LN}(\mathbf{U}_v))), \end{aligned} \tag{9}$$

where $\mathbf{U}_v$ denotes the intermediate representation after the attention residual connection. The audio-conditioned textual representation $\mathbf{A}_a$ is obtained analogously using $\mathbf{H}_a$ as the source representation. Since both cross-modal branches use the textual sequence as the query and operate in the common hidden dimension D, $\mathbf{A}_a$, $\mathbf{A}_v$, and $\mathbf{H}_t$ have compatible temporal and feature dimensions.

Following the text-guided fusion scheme, the textual representation and the two cross-modal representations are first integrated as

$$\mathbf{Z}_m = \mathbf{H}_t + \mathbf{A}_a + \mathbf{A}_v. \tag{10}$$

A self-attention transformation is subsequently applied to further capture interactions within the integrated multimodal sequence:

$$\mathbf{G}_m = \mathrm{SelfAttn}(\mathbf{Z}_m). \tag{11}$$

In parallel, the learned MLD $\mathbf{d} = [d_t, d_a, d_v]$ modulates the original modality-specific representations as

$$\widehat{\mathbf{H}}_t = d_t \mathbf{H}_t, \qquad \widehat{\mathbf{H}}_a = d_a \mathbf{H}_a, \qquad \widehat{\mathbf{H}}_v = d_v \mathbf{H}_v. \tag{12}$$

The text-guided multimodal representation and the MLD-modulated unimodal representations are then concatenated and projected to obtain the fused sequence:

$$\mathbf{F} = \mathbf{W}_{\mathrm{fus}}\left[\mathbf{G}_m; \widehat{\mathbf{H}}_t; \widehat{\mathbf{H}}_a; \widehat{\mathbf{H}}_v\right] + \mathbf{b}_{\mathrm{fus}}, \tag{13}$$

where $[\cdot;\cdot]$ denotes feature-wise concatenation, and $\mathbf{W}_{\mathrm{fus}}$ and $\mathbf{b}_{\mathrm{fus}}$ denote the learnable projection parameters of the fusion layer. The resulting fused representation $\mathbf{F}$ is subsequently passed to the MLD-Informed Sparse Expert Routing module.

3.4. MLD-Informed Sparse Expert Routing Module

Sparsely gated Mixture-of-Experts (MoE) models [21] employ a learnable router to activate a subset of experts for each input. In many existing multimodal MoE methods, expert selection is primarily conditioned on fused semantic representations, while sample-dependent modality weighting information is not explicitly incorporated into the routing process. To address this limitation, the proposed MLD-Informed Sparse Expert Routing module conditions expert selection jointly on the fused multimodal representation and the learned MLD.

Given the MLD-modulated fused representation $\mathbf{F}$ obtained from the preceding subsection, the final temporal representation of the fused sequence is used as the sample-level semantic representation, i.e., $\mathbf{f}_p = \mathbf{F}_{-1}$. The semantic representation $\mathbf{f}_p$ and the learned MLD $\mathbf{d} = [d_t, d_a, d_v]$ are independently mapped to expert-level routing logits.

Specifically, the feature-derived and MLD-derived routing logits are computed as

$$\boldsymbol{\ell}^F = \mathbf{W}_F \mathbf{f}_p, \qquad \boldsymbol{\ell}^d = \mathbf{W}_d \mathbf{d}, \tag{14}$$

where $\mathbf{W}_F$ and $\mathbf{W}_d$ are learnable mappings from the fused semantic representation and the MLD, respectively, to the n_e-dimensional expert-logit space.

The two routing components are combined additively to obtain the clean routing logits:

$$\boldsymbol{\ell} = \boldsymbol{\ell}^F + \boldsymbol{\ell}^d + \mathbf{b}, \tag{15}$$

where $\mathbf{b} \in \mathbb{R}^{n_e}$ is a learnable expert-wise bias. This direct additive formulation allows the fused multimodal semantics and the MLD to contribute complementary evidence to expert selection without introducing an additional hidden routing representation.

During training, Gaussian perturbation is introduced before sparse expert selection to encourage routing exploration. The expert-wise noise scale is predicted from the fused semantic representation as

$$\boldsymbol{\sigma} = \mathrm{Softplus}\left(\mathbf{W}_\sigma \mathbf{f}_p + \mathbf{b}_\sigma\right), \tag{16}$$

and the routing logits used for expert selection are

$$\tilde{\boldsymbol{\ell}} = \boldsymbol{\ell} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{17}$$

The Gaussian perturbation is applied only during training, whereas the clean logits $\boldsymbol{\ell}$ are used during evaluation. Let $\mathcal{K}(\tilde{\boldsymbol{\ell}})$ denote the index set of the k experts with the largest routing logits. The sparse routing weights are obtained by applying softmax normalization over the selected experts:

$$g_i = \begin{cases} \dfrac{\exp(\tilde{\ell}_i)}{\sum_{j \in \mathcal{K}(\tilde{\boldsymbol{\ell}})} \exp(\tilde{\ell}_j)}, & i \in \mathcal{K}(\tilde{\boldsymbol{\ell}}), \\ 0, & i \notin \mathcal{K}(\tilde{\boldsymbol{\ell}}). \end{cases} \tag{18}$$

Softmax normalization is performed only over the selected Top-k experts. During backpropagation, gradients propagate through the selected routing weights and activated expert branches, whereas experts that are not selected for a given sample receive no sample-specific expert gradient from that instance. Each expert is implemented as a feed-forward network with the same architecture but independent parameters. The selected expert outputs are weighted and aggregated as

$$\mathbf{z} = \sum_{i \in \mathcal{K}(\tilde{\boldsymbol{\ell}})} g_i E_i(\mathbf{f}_p), \tag{19}$$

where $E_i(\cdot)$ denotes the i-th expert. By combining semantic and MLD-derived routing evidence before sparse Top-k selection, the module produces an MLD-informed expert representation $\mathbf{z}$ for subsequent sentiment prediction.

3.5. Prediction and Training Objective

The MLD-informed expert representation $\mathbf{z}$ is passed through a two-layer prediction MLP, consisting of a hidden linear layer followed by ReLU activation and dropout and a final linear regression layer, to produce the sentiment prediction:

$$\hat{y} = \mathrm{MLP}(\mathbf{z}), \tag{20}$$

For a mini-batch of B samples, the training objective is defined using the mean absolute error (MAE) loss:

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{B} \sum_{n=1}^{B} |y_n - \hat{y}_n|, \tag{21}$$

where y_n and $\hat{y}_n$ denote the ground-truth and predicted sentiment scores of the n-th sample, respectively.

Since explicit MLD annotations are unavailable, no auxiliary supervision is imposed on the description degrees. No auxiliary expert load-balancing loss is introduced in the current formulation; expert selection is learned solely through the downstream sentiment objective. The entire framework is jointly optimized end-to-end using $\mathcal{L}_{\mathrm{MAE}}$. Accordingly, the MLD is learned indirectly from the sentiment prediction objective and is subsequently used to modulate multimodal fusion and inform sparse expert routing.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate the proposed MLD-FER framework on two widely used multimodal sentiment analysis datasets, CMU-MOSI [22] and CMU-MOSEI [23]. CMU-MOSI contains 2199 opinion video segments collected from 93 YouTube videos. The original CMU-MOSEI dataset contains 23,453 annotated video segments collected from 3228 videos involving 1000 speakers. Each segment in both datasets is annotated with a sentiment score ranging from −3 to +3, where larger values indicate more positive sentiment. Each segment contains textual, acoustic, and visual modalities. Following the standard data split, using the word-aligned

processed data distributed through MultiBench [24], CMU-MOSI contains 1284 training, 229 validation, and 686 test samples. For CMU-MOSEI, the corresponding processed version contains 22,856 samples, including 16,326 training, 1871 validation, and 4659 test samples.

Evaluation Metrics. Following the evaluation protocols adopted in prior MSA studies [6,9,10], we assess the continuous sentiment predictions using both regression and classification-based metrics. For regression, we report mean absolute error (MAE) and Pearson correlation coefficient (Corr). A lower MAE indicates better prediction accuracy, while a higher Corr indicates stronger linear agreement between the predicted and ground-truth sentiment scores. For classification-based evaluation, we report seven-class accuracy (Acc-7), binary accuracy (Acc-2), and F1 score. Acc-7 measures the accuracy of predictions after mapping the continuous sentiment scores to seven discrete levels ranging from −3 to +3. For Acc-2 and F1, following TGMoE [9], two evaluation settings are considered: negative versus non-negative, where positive and neutral samples are grouped together, and negative versus positive, where neutral samples are excluded.

4.2. Experimental Settings

The proposed MLD-FER model is implemented in PyTorch (version 2.4.1), and all experiments are conducted on an NVIDIA RTX 3090 GPU. The experiments use the word-aligned multimodal data provided by MultiBench [24], with a common sequence length of 50. The textual input is further encoded and jointly fine-tuned using BERT-base-uncased, while the acoustic and visual inputs directly use the pre-extracted COVAREP and Facet features provided by MultiBench, respectively. The main implementation settings are summarized in Table 1.

Table 1. Experimental settings on the CMU-MOSI and CMU-MOSEI datasets.

Setting	CMU-MOSI	CMU-MOSEI
Text feature dimension	768	768
Acoustic feature dimension	5	74
Visual feature dimension	20	35
Sequence length	50	50
Common feature dimension	768	768
Batch size	32	32
Main learning rate	1×10^{-4}	1×10^{-3}
BERT learning rate	1×10^{-4}	5×10^{-5}
Weight decay	0.005	0.0005
Maximum epochs	50	50
Cross-modal Transformer layers	4	4
Attention heads	8	8
MLD scoring MLP hidden dimension	32	32
MLD scoring activation	ReLU	ReLU
Expert hidden dimension	128	128
Expert activation	ReLU	ReLU
Number of experts	3	3
Top-K experts	2	2
Dropout rate	0.1	0.1
Number of parameters	197.72M	197.78M

The model is optimized using the Adam optimizer. The final hyperparameter configuration is selected according to validation-set performance. An early stopping strategy is adopted, with training terminated when the validation MAE does not improve for 10 consecutive epochs. To account for variability due to random initialization, the reported results of MLD-FER are obtained over five random seeds: 1111, 1112, 1113, 1114, and 1115.

4.3. Comparison with Existing Methods

To evaluate the overall performance of MLD-FER, comparisons are conducted with representative multimodal sentiment analysis methods on the CMU-MOSI and CMU-MOSEI datasets. The comparison results are reported in Tables 2 and 3, respectively. For Acc-2 and F1, the values on the left and right sides of "/" correspond to the negative/non-negative and negative/positive settings, respectively. The results of the conventional baseline models are taken from TGMoE [9], while the results of the additional recent methods are obtained from their corresponding original publications.

Table 2. Comparison results on the CMU-MOSI dataset.

Model	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 ↑	F1 ↑
TFN [25]	0.901	0.698	34.90	–/80.80	–/80.70
LMF [26]	0.917	0.695	33.20	–/82.50	–/82.40
MFM [27]	0.877	0.706	35.40	–/81.70	–/81.60
MulT [4]	0.861	0.711	–	81.50/84.10	80.60/83.90
MAG-BERT [5]	0.727	0.781	43.62	82.37/84.43	82.50/84.61
MISA [6]	0.804	0.764	–	80.79/82.10	80.77/82.03
BIMHA [28]	0.925	0.671	36.44	78.57/80.30	78.57/80.30
SELF-MM [7]	0.712	0.795	45.79	82.54/84.77	82.68/84.91
CubeMLP [29]	0.770	0.767	45.50	–/85.60	–/85.50
TETFN [8]	0.717	0.800	–	84.05/86.10	83.83/86.07
MISA-GMC [18]	0.712	0.795	45.92	82.94/85.21	82.75/85.10
EMOE [10]	0.710	–	47.70	–/85.40	–/85.40
TGMoE [9]	0.760	0.767	45.89	**85.37**/85.64	**85.23**/85.71
MLD-FER (ours)	$\mathbf{0.701}_{0.007}$	$\mathbf{0.807}_{0.008}$	$\mathbf{48.40}_{0.36}$	$85.26_{0.28}$/$\mathbf{86.28}_{0.31}$	$85.13_{0.26}$/$\mathbf{86.18}_{0.28}$

Note: ↓ and ↑ indicate lower-is-better and higher-is-better metrics, respectively. Bold values indicate the best results for each metric. Subscripts denote standard deviations over five random seeds. "–" indicates that the corresponding result was not reported.

Table 3. Comparison results on the CMU-MOSEI dataset.

Model	MAE ↓	Corr ↑	Acc-7 ↑	Acc-2 ↑	F1 ↑
TFN [25]	0.593	0.700	50.20	–/82.50	–/82.10
LMF [26]	0.623	0.677	48.00	–/82.00	–/82.10
MFM [27]	0.568	0.717	51.30	–/84.40	–/84.30
MulT [4]	0.580	0.703	–	–/82.50	–/82.30
MAG-BERT [5]	0.543	0.755	52.67	82.51/84.82	82.77/84.71
MISA [6]	0.568	0.724	–	82.59/84.23	82.67/83.97
BIMHA [28]	0.559	0.731	52.11	84.07/83.96	83.35/83.50
SELF-MM [7]	**0.529**	0.767	53.46	82.68/84.96	82.95/84.93
CubeMLP [29]	**0.529**	0.760	**54.90**	–/85.10	–/84.50
TETFN [8]	0.551	0.748	–	84.25/85.18	84.18/85.27
MISA-GMC [18]	0.560	0.756	51.62	84.18/85.50	84.25/85.27
EMOE [10]	0.536	–	54.10	–/85.30	–/85.30
TGMoE [9]	0.535	0.757	53.70	**85.36**/85.51	**85.58**/85.86
MLD-FER (ours)	$0.531_{0.004}$	$\mathbf{0.768}_{0.002}$	$54.17_{0.35}$	$85.34_{0.17}$/$\mathbf{86.13}_{0.15}$	$85.50_{0.11}$/$\mathbf{86.01}_{0.18}$

Note: ↓ and ↑ indicate lower-is-better and higher-is-better metrics, respectively. Bold values indicate the best results for each metric. Subscripts denote standard deviations over five random seeds. "–" indicates that the corresponding result was not reported.

Performance on the CMU-MOSI dataset. As shown in Table 2, MLD-FER achieves strong overall performance on the CMU-MOSI dataset. For regression, MLD-FER achieves the lowest MAE among the compared methods, outperforming EMOE [10] (0.701 vs. 0.710) with an absolute reduction of 0.009; compared with the direct baseline TGMoE [9], the

reduction is 0.059 (0.701 vs. 0.760). It also obtains the highest Corr, exceeding TETFN [8] (0.807 vs. 0.800) by 0.007 and TGMoE (0.807 vs. 0.767) by 0.040. For seven-class classification, MLD-FER achieves the highest Acc-7 of 48.40%, outperforming EMOE (48.40% vs. 47.70%) by 0.70 percentage points and TGMoE (48.40% vs. 45.89%) by 2.51 percentage points. Under the negative/positive setting, MLD-FER also achieves the best Acc-2 and F1 scores, exceeding TETFN in Acc-2 (86.28% vs. 86.10%) and F1 (86.18% vs. 86.07%) by 0.18 and 0.11 percentage points, respectively; relative to TGMoE, the corresponding improvements are 0.64 and 0.47 percentage points. Under the negative/non-negative setting, TGMoE remains slightly better, with Acc-2 of 85.37% versus 85.26% and F1 of 85.23% versus 85.13%, corresponding to differences of only 0.11 and 0.10 percentage points. Overall, MLD-FER shows particularly clear improvements in regression, seven-class classification, and negative/positive sentiment discrimination while remaining competitive under the negative/non-negative setting.

Performance on the CMU-MOSEI dataset. As shown in Table 3, MLD-FER achieves competitive overall performance on the CMU-MOSEI dataset. For regression, MLD-FER obtains an MAE of 0.531, which is only 0.002 higher than the lowest reported value of 0.529 achieved by SELF-MM [7] and CubeMLP [29], while reducing the MAE of TGMoE [9] by 0.004 (0.531 vs. 0.535). In terms of Corr, MLD-FER achieves the highest value of 0.768, exceeding SELF-MM (0.768 vs. 0.767) by 0.001 and TGMoE (0.768 vs. 0.757) by 0.011. For seven-class classification, MLD-FER achieves an Acc-7 of 54.17%, which is lower than CubeMLP (54.17% vs. 54.90%) by 0.73 percentage points but exceeds TGMoE (54.17% vs. 53.70%) by 0.47 percentage points and EMOE [10] (54.17% vs. 54.10%) by 0.07 percentage points. Under the negative/positive setting, MLD-FER achieves the highest Acc-2 and F1 scores, outperforming TGMoE in Acc-2 (86.13% vs. 85.51%) and F1 (86.01% vs. 85.86%) by 0.62 and 0.15 percentage points, respectively. Under the negative/non-negative setting, MLD-FER remains comparable to TGMoE, with differences of only 0.02 percentage points in Acc-2 (85.34% vs. 85.36%) and 0.08 percentage points in F1 (85.50% vs. 85.58%). Overall, MLD-FER maintains a favorable balance between regression and classification performance on CMU-MOSEI, achieving the best Corr and negative/positive classification results while remaining competitive on the other metrics.

Taken together, the results on both datasets show that MLD-FER consistently achieves competitive performance across regression, fine-grained classification, and binary sentiment evaluation. In particular, the improvements under the negative/positive protocol suggest that incorporating sample-dependent MLD weights into cross-modal fusion and expert routing can provide useful task-adaptive information for sentiment prediction. The contributions of these two uses of MLD, as well as the statistical descriptors from which the MLD is estimated, are examined further in the following ablation studies.

4.4. Ablation Study

We conduct a series of ablation and diagnostic experiments to examine the effects of the proposed designs in MLD-FER. We first investigate the use of the learned MLD in cross-modal fusion and expert routing through controlled component ablations. We then examine the statistical descriptors used for MLD estimation through correlation analysis and leave-one-descriptor-out experiments.

4.4.1. Component Ablation

We conduct controlled component ablations on CMU-MOSI and CMU-MOSEI to examine the effects of using the learned MLD in cross-modal fusion and expert routing, while keeping all other architectural and training settings unchanged. When MLD-modulated fusion is disabled, the learned modality weights are replaced with uniform weights, $d_t = d_a = d_v = 1/3$, while the fusion structure is retained. When MLD-informed

routing is disabled, the MLD-dependent routing term is removed, leaving expert scores determined solely by the fused sample representation. MLD-FER retains the text-guided cross-modal interaction and sparse-expert processing paradigm of TGMoE [9], while introducing the statistical-descriptor-based MLD estimator, MLD-modulated fusion, and MLD-informed routing mechanisms. Therefore, the ×/× variant should not be interpreted as an exact reproduction of TGMoE, which also explains its small numerical difference from the originally reported TGMoE results in Tables 2 and 3.

As shown in Table 4a, removing either use of MLD consistently degrades the overall performance on CMU-MOSI. When MLD-informed routing is removed, MAE increases from 0.701 to 0.735, Corr decreases from 0.807 to 0.775, and Acc-7 decreases by 2.34 percentage points (48.40% vs. 46.06%). The negative/positive Acc-2 and F1 scores also decrease by 0.59 and 0.47 percentage points, respectively. Removing MLD-modulated fusion results in a slightly smaller but still consistent degradation, with MAE increasing to 0.729, Corr decreasing to 0.781, and Acc-7 decreasing by 2.06 percentage points to 46.34%. Under the negative/positive setting, Acc-2 and F1 decrease by 0.56 and 0.43 percentage points, respectively. When both uses of MLD are removed, the degradation becomes more pronounced: MAE increases to 0.751, Corr decreases to 0.755, and Acc-7 decreases by 2.53 percentage points to 45.87%. The negative/positive Acc-2 and F1 scores also decrease by 0.98 and 0.85 percentage points, respectively.

Table 4. Component ablation results on the CMU-MOSI and CMU-MOSEI datasets.

(a) Results on CMU-MOSI.						
MLD Usage		**Performance**				
Fusion	**Routing**	**MAE ↓**	**Corr ↑**	**Acc-7 ↑**	**Acc-2 ↑**	**F1 ↑**
×	×	0.751	0.755	45.87	84.91/85.30	84.92/85.33
✓	×	0.735	0.775	46.06	85.00/85.69	84.98/85.71
×	✓	0.729	0.781	46.34	85.13/85.72	**85.13**/85.75
✓	✓	**0.701**	**0.807**	**48.40**	**85.26/86.28**	**85.13/86.18**
(b) Results on CMU-MOSEI.						
MLD Usage		**Performance**				
Fusion	**Routing**	**MAE ↓**	**Corr ↑**	**Acc-7 ↑**	**Acc-2 ↑**	**F1 ↑**
×	×	0.537	0.754	53.54	85.32/85.50	85.38/85.63
✓	×	0.535	0.759	53.72	**85.34**/85.74	85.37/85.75
×	✓	0.533	0.762	53.85	**85.34**/85.85	85.40/85.86
✓	✓	**0.531**	**0.768**	**54.17**	**85.34/86.13**	**85.50/86.01**

Note: ↓ and ↑ indicate lower-is-better and higher-is-better metrics, respectively. ✓ and × indicate enabled and disabled components, respectively. Bold values indicate the best results for each metric.

A similar pattern is observed on CMU-MOSEI, as shown in Table 4b. Removing MLD-informed routing increases MAE from 0.531 to 0.535, reduces Corr from 0.768 to 0.759, and decreases Acc-7 by 0.45 percentage points. Removing MLD-modulated fusion produces smaller decreases, with an MAE of 0.533, a Corr of 0.762, and an Acc-7 of 53.85%, corresponding to a decrease of 0.32 percentage points. Under the negative/positive setting, removing routing decreases Acc-2 and F1 by 0.39 and 0.26 percentage points, whereas removing fusion decreases them by 0.28 and 0.15 percentage points, respectively. When both components are removed, MAE further increases to 0.537, Corr decreases to 0.754, and Acc-7 decreases by 0.63 percentage points to 53.54%. The negative/positive Acc-2 and F1 scores decrease by 0.63 and 0.38 percentage points, respectively.

Overall, the controlled ablations demonstrate that MLD-modulated fusion and MLD-informed routing provide complementary benefits. Removing either component degrades

regression, correlation, and fine-grained classification performance on both datasets, while jointly removing them produces the largest overall degradation. Notably, removing MLD-informed routing leads to slightly larger decreases than removing MLD-modulated fusion for MAE, Corr, Acc-7, and most negative/positive classification metrics on both datasets. This suggests that conditioning the sparse router on the learned MLD provides useful sample-dependent information for expert selection, while MLD-modulated fusion contributes additional gains by adapting the relative weighting of the three modality branches before routing. The effect is more evident under the negative/positive protocol than under the negative/non-negative protocol, particularly on CMU-MOSEI, where the latter metric is relatively insensitive to the individual ablations. These results indicate that the two mechanisms operate at different stages of the model and are not functionally interchangeable; using the same learned MLD for both representation fusion and expert routing yields the most consistent overall performance.

4.4.2. Statistical Descriptor Analysis

The proposed MLD is estimated from three modality-level statistical descriptors, namely standard deviation (SD), peak-to-dispersion ratio (PDR), and zero-rate (ZR). To examine their relationships and their contributions to MLD estimation, we conduct two complementary analyses. First, we measure the pairwise Pearson correlations among the three descriptors for each modality to characterize their statistical relationships. Second, we perform leave-one-descriptor-out ablations by removing one descriptor at a time while keeping the model architecture and training settings unchanged. For each ablation, the removed descriptor is set to zero after normalization using statistics estimated from the training set, while the dimensionality and structure of the shared MLD scoring network remain unchanged.

Table 5 shows that the relationships among the three statistical descriptors are modality dependent. For the textual modality, SD and PDR exhibit strong negative correlations on both CMU-MOSI ($\rho = -0.863$) and CMU-MOSEI ($\rho = -0.781$), reflecting the mathematical relationship between dispersion and the PDR formulation. In contrast, the SD–PDR correlations are substantially weaker for the acoustic modality on CMU-MOSI ($\rho = -0.294$) and CMU-MOSEI ($\rho = -0.116$), while the visual modality exhibits moderate correlations on CMU-MOSI ($\rho = -0.324$) and CMU-MOSEI ($\rho = 0.346$). The correlations involving ZR are generally weak to moderate for the acoustic and visual modalities, with the largest absolute value being the SD–ZR correlation for the visual modality on CMU-MOSEI ($\rho = -0.568$). After padded temporal positions are excluded, textual ZR is identically zero and therefore has zero variance, making its Pearson correlations with SD and PDR undefined, as indicated by "–" in the table. Overall, the results indicate that the statistical relationships among SD, PDR, and ZR vary substantially across modalities rather than exhibiting a uniform redundancy pattern.

Table 5. Pairwise Pearson correlations among the statistical descriptors on CMU-MOSI and CMU-MOSEI.

Dataset	Modality	ρ(SD, PDR)	ρ(SD, ZR)	ρ(PDR, ZR)
CMU-MOSI	Text	−0.863	–	–
	Audio	−0.294	−0.374	0.061
	Vision	−0.324	−0.369	0.182
CMU-MOSEI	Text	−0.781	–	–
	Audio	−0.116	0.070	−0.231
	Vision	0.346	−0.568	−0.354

Note: "–" indicates that Pearson's correlation is undefined because textual ZR has zero variance after padded positions are excluded.

The leave-one-descriptor-out results in Table 6 further reveal different effects of the three descriptors. On CMU-MOSI, removing SD produces the largest degradation in regression and fine-grained classification, increasing MAE from 0.701 to 0.749 and reducing Acc-7 from 48.40% to 45.91%. Removing PDR has a smaller effect on MAE and Acc-7 but produces the largest decrease in binary classification performance, with Acc-2 decreasing from 85.26/86.28% to 84.92/85.31% and F1 decreasing from 85.13/86.18% to 84.92/85.32%. Removing ZR results in the smallest overall degradation, although the full descriptor set still yields better Corr, Acc-7, and binary classification results.

Table 6. Leave-one-descriptor-out ablation results on the CMU-MOSI and CMU-MOSEI datasets.

(a) Results on CMU-MOSI.								
Variant	**SD**	**PDR**	**ZR**	**MAE ↓**	**Corr ↑**	**Acc-7 ↑**	**Acc-2 ↑**	**F1 ↑**
w/o SD	×	✓	✓	0.749	0.784	45.91	85.01/85.54	85.03/85.55
w/o PDR	✓	×	✓	0.731	0.787	46.65	84.92/85.31	84.92/85.32
w/o ZR	✓	✓	×	0.713	0.798	46.79	85.14/85.85	85.12/85.87
Full	✓	✓	✓	**0.701**	**0.807**	**48.40**	**85.26/86.28**	**85.13/86.18**
(b) Results on CMU-MOSEI.								
Variant	**SD**	**PDR**	**ZR**	**MAE ↓**	**Corr ↑**	**Acc-7 ↑**	**Acc-2 ↑**	**F1 ↑**
w/o SD	×	✓	✓	0.537	0.756	53.55	85.33/85.63	85.39/85.69
w/o PDR	✓	×	✓	0.535	0.760	53.78	85.32/85.54	85.37/85.61
w/o ZR	✓	✓	×	**0.531**	0.762	53.85	**85.34**/85.89	85.43/85.91
Full	✓	✓	✓	**0.531**	**0.768**	**54.17**	**85.34/86.13**	**85.50/86.01**

Note: ↓ and ↑ indicate lower-is-better and higher-is-better metrics, respectively. ✓ and × indicate enabled and disabled components, respectively. Bold values indicate the best results for each metric.

A similar tendency is observed on CMU-MOSEI. Removing SD increases MAE from 0.531 to 0.537 and reduces Corr from 0.768 to 0.756, representing the largest regression degradation among the three leave-one-out variants. Removing PDR again produces the largest decrease in binary classification performance, particularly under the negative/positive setting, where Acc-2 decreases from 86.13% to 85.54% and F1 decreases from 86.01% to 85.61%. In comparison, removing ZR has a smaller overall effect: MAE remains 0.531, while Corr and Acc-7 decrease from 0.768 and 54.17% to 0.762 and 53.85%, respectively.

Taken together with the correlation analysis, these results indicate that statistical correlation does not necessarily imply functional redundancy in the learned MLD. Although SD and PDR are strongly correlated for the textual modality, removing them leads to different downstream degradation patterns: SD has a larger effect on regression and fine-grained sentiment prediction, whereas PDR has a larger effect on binary sentiment discrimination. ZR contributes more modestly overall, which is consistent with the observation that textual ZR becomes constant after padded positions are excluded, leaving its sample-dependent information primarily in the acoustic and visual modalities. The full descriptor set nevertheless provides the most consistent overall performance across the two datasets and evaluation protocols.

4.5. Visualization Analysis

We further examine the behavior of the learned MLD and sparse expert routing through test-set visualizations. We first visualize the distribution of the learned MLD over all test samples to examine its sample-dependent weighting patterns. We then analyze expert utilization and perform forced-routing evaluation on CMU-MOSI to examine both expert participation and the predictive relevance of the learned expert ranking.

4.5.1. Distribution of the Learned MLD

To examine the sample-level behavior of the learned MLD, we visualize the MLD of every test sample on CMU-MOSI and CMU-MOSEI using ternary plots. Each point represents one test sample and is determined by its three learned modality weights, $\mathbf{d} = [d_t, d_a, d_v]$, where $d_t + d_a + d_v = 1$. The center of the simplex corresponds to uniform modality weighting, i.e., $(1/3, 1/3, 1/3)$.

As illustrated in Figure 2, the learned MLDs on both datasets occupy a relatively balanced central region of the simplex while exhibiting continuous sample-level variation around it. Importantly, the distributions do not collapse to a single uniform point $(1/3, 1/3, 1/3)$; instead, individual samples are assigned different modality-weighting patterns according to their modality-level statistical descriptors. Some samples remain relatively balanced, whereas others show more pronounced shifts toward particular modality directions, as reflected by the peripheral samples. This distribution pattern suggests that MLD-FER does not rely on globally fixed modality weights or force highly polarized assignments. Rather, it performs moderate, sample-dependent reweighting of the three modalities, allowing the weighting pattern to adapt continuously across samples.

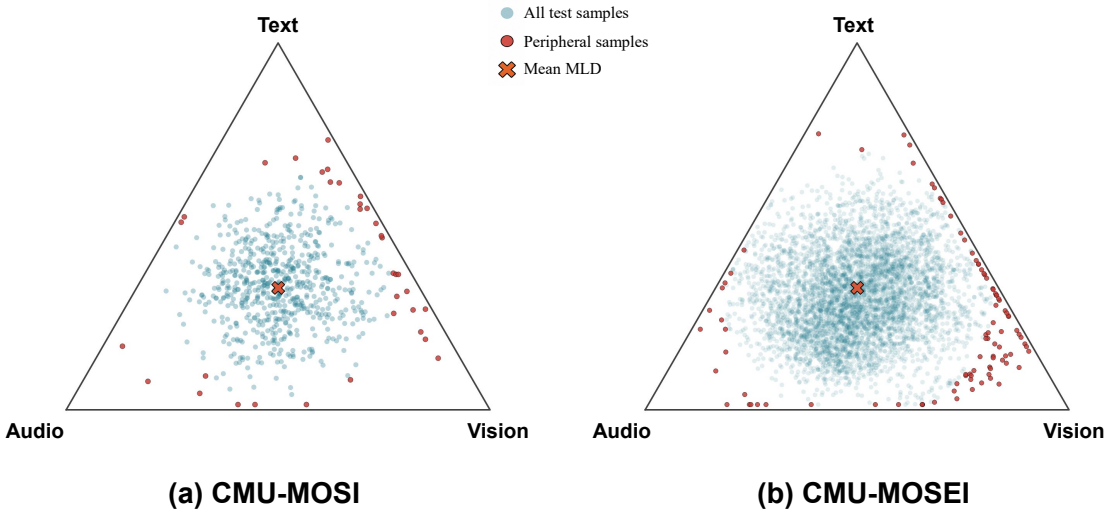


Figure 2. Visualization of the learned MLDs on the test sets of (**a**) CMU-MOSI and (**b**) CMU-MOSEI. Each point represents the MLD of one test sample, defined by the three modality weights (d_t, d_a, d_v) with $d_t + d_a + d_v = 1$. The orange cross denotes the mean MLD, while the red points indicate peripheral samples with relatively large deviations from the mean MLD.

4.5.2. Expert Utilization and Routing Validation

We further examine the sparse expert router from two complementary perspectives: expert utilization and routing effectiveness. Since MLD-FER does not employ an auxiliary load-balancing loss, Figure 3a reports the test-set selection frequency of each expert across three independently trained checkpoints. Figure 3b further evaluates whether the router's expert ranking is consistent with actual predictive utility through forced-routing experiments.

As shown in Figure 3a, all three experts remain active across the evaluated seeds, although the degree of utilization imbalance varies with random initialization. Seeds 1111 and 1113 exhibit relatively distributed usage among the three experts, whereas Seed 1112 shows a more uneven allocation, with E3 selected for all test samples and E2 selected less frequently. Nevertheless, no expert is completely inactive in any of the three runs.

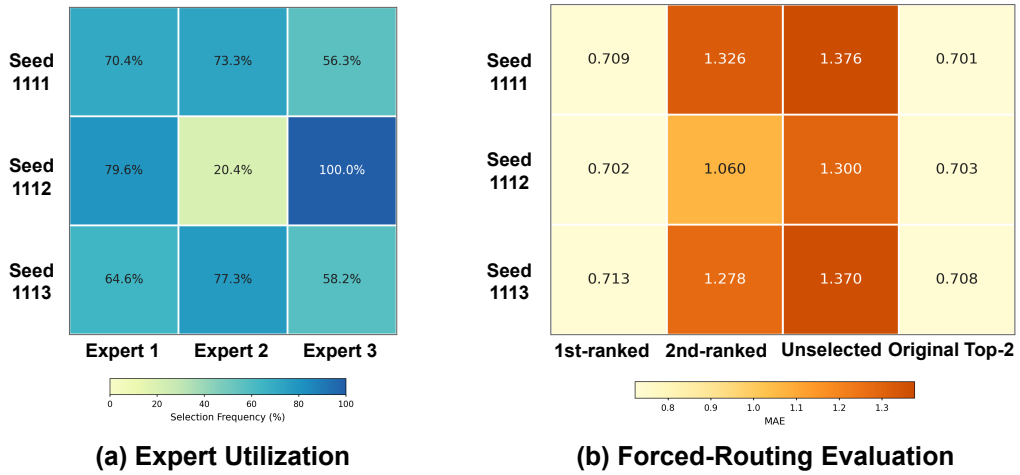


Figure 3. Expert utilization and forced-routing evaluation on the CMU-MOSI test set across three random seeds. (**a**) Test-set selection frequency of each expert under Top-2 routing; because two experts are activated for each sample, the three frequencies for each seed sum to 200%. (**b**) MAE obtained by retaining only the first-ranked expert, the second-ranked expert, or the unselected expert for each sample, together with the original Top-2 weighted routing. Lower MAE indicates better predictive performance.

For the forced-routing analysis in Figure 3b, the two selected experts of each test sample are ordered by their routing weights. The 1st-ranked condition retains only the higher-weight expert, the 2nd-ranked condition retains only the other selected expert, and the Unselected condition forces prediction through the expert excluded by Top-2 routing. These expert identities are determined separately for each sample and are therefore not tied to fixed indices such as E1 or E2. Across all three seeds, the first-ranked expert consistently yields lower MAE than the second-ranked expert, which in turn outperforms the unselected expert. The original Top-2 aggregation achieves performance comparable to or slightly better than using only the first-ranked expert. These results indicate that the learned routing scores are meaningfully associated with sample-specific expert utility, while all experts remain involved in sparse prediction.

5. Conclusions

This work investigates sample-dependent modality weighting in multimodal sentiment analysis from a statistical perspective and further explores how such weighting information can be used to coordinate multimodal fusion and sparse expert routing. To this end, we propose Multimodal Label Distribution-Guided Fusion and Expert Routing (MLD-FER). MLD-FER first learns a normalized, sample-dependent Multimodal Label Distribution (MLD) from modality-level statistical descriptors. The learned MLD serves as a task-adaptive, sample-dependent modality weighting signal that modulates multimodal fusion and provides additional conditioning information for sparse expert routing. In this way, the same MLD coordinates cross-modal fusion and sparse expert routing as a shared sample-dependent signal.

Experiments on CMU-MOSI and CMU-MOSEI show that MLD-FER achieves competitive performance across regression and classification metrics. The component ablations show that MLD-modulated fusion and MLD-informed routing provide complementary contributions, while the descriptor correlation and leave-one-descriptor-out analyses further examine the contributions of the statistical descriptors in MLD estimation. The test-set visualization shows that the learned MLD exhibits continuous sample-dependent variation around a relatively balanced region rather than collapsing to a fixed weighting pattern.

In addition, expert-utilization analysis confirms that all experts participate in prediction across the examined random seeds, and the forced-routing evaluation consistently shows lower prediction error for higher-ranked experts than for lower-ranked or unselected experts. These results collectively indicate that the proposed statistical side-information mechanism provides useful adaptive signals for both multimodal representation learning and sparse expert routing.

Future work will investigate richer statistical and learnable side information for MLD estimation and explore whether the learned MLD can serve as a sample-dependent coordination signal at other stages of multimodal representation learning, such as cross-modal interaction, modality alignment, and representation refinement. More explicit expert coordination mechanisms may also be explored to further improve routing stability and expert diversity. In addition, extending MLD-FER to broader multimodal affective computing and representation learning tasks, including more challenging settings with noisy or incomplete modalities, will help further assess its generalizability and practical applicability.

Author Contributions: Conceptualization, L.Y.; methodology, L.Y.; software, L.Y.; validation, L.Y.; investigation, L.Y.; resources, H.S.; data curation, L.Y.; writing—original draft preparation, L.Y.; writing—review and editing, H.S.; visualization, L.Y.; funding acquisition, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI Grant Number JP26K14937.

Data Availability Statement: No new datasets were created in this study. The original contributions presented in this study are included in the article. The CMU-MOSI and CMU-MOSEI datasets analyzed in this study are publicly available. Information on accessing these datasets is provided in the corresponding original publications cited in this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, H.; Zhao, Y.; Wu, Y.; Wang, S.; Zheng, T.; Zhang, H.; Ma, Z.; Che, W.; Wang, S.; Wei, S.; et al. Large language models meet text-centric multimodal sentiment analysis: A survey. *Sci. China Inf. Sci.* **2025**, *68*, 200101. [CrossRef]
2. Dhotay, M.; Dharrao, M.; Deokate, S.; Bongale, A.; Dharrao, D. Multimodal sentiment analysis: Emerging innovations, core challenges, and future directions. *Discov. Artif. Intell.* **2026**, *6*, 433. [CrossRef]
3. Zhao, K.; Zheng, M.; Li, Q.; Liu, J. Multimodal sentiment analysis—A comprehensive survey from a fusion methods perspective. *IEEE Access* **2025**, *13*, 64556–64583. [CrossRef]
4. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6558–6569. [CrossRef]
5. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.B.; Mao, C.; Morency, L.-P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 2359–2369. [CrossRef]
6. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: Modality-invariant and modality-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131. [CrossRef]
7. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 10790–10797. . [CrossRef]
8. Wang, D.; Guo, X.; Tian, Y.; Liu, J.; He, L.; Luo, X. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognit.* **2023**, *136*, 109259. [CrossRef]
9. Zhao, X.; Wang, M.; Tan, Y.; Wang, X. TGMoE: A text guided mixture-of-experts model for multimodal sentiment analysis. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 1227. [CrossRef]
10. Fang, Y.; Huang, W.; Wan, G.; Su, K.; Ye, M. EMOE: Modality-specific enhanced dynamic emotion experts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 11–15 June 2025; pp. 14314–14324. [CrossRef]

11. Ren, Y.; Xu, N.; Ling, M.; Geng, X. Label distribution for multimodal machine learning. *Front. Comput. Sci.* **2022**, *16*, 161306. [CrossRef]
12. Song, X.; Liu, C.; Xu, L.; Gao, B.; Lu, Z.; Zhang, Y. Affective computing methods for multimodal embodied AI human–computer interaction. *Aslib J. Inf. Manag.* **2025**, 1–25. [CrossRef]
13. Guerrero-Sosa, J.D.T.; Romero, F.P.; Menéndez-Domínguez, V.H.; Serrano-Guerrero, J.; Montoro-Montarroso, A.; Olivas, J.A. A comprehensive review of multimodal analysis in education. *Appl. Sci.* **2025**, *15*, 5896. [CrossRef]
14. Ghosh, A.; Umer, S.; Dhara, B.C.; Ali, G.G.M.N. A multimodal pain sentiment analysis system using ensembled deep learning approaches for IoT-enabled healthcare framework. *Sensors* **2025**, *25*, 1223. [CrossRef]
15. Baberwal, S.K.; Shelke, N.A.; Anwar, K. Systematic review of recent advances in multimodal sentiment analysis. *Discov. Comput.* **2025**, *28*, 270. [CrossRef]
16. Anwar, K.; Shreya; Sharma, M.; Saanvi, K. A review of multimodal sentiment analysis: Taxonomy, issues, challenges, and future perspectives. *Comput. Electr. Eng.* **2026**, *131*, 110959. [CrossRef]
17. Han, W.; Chen, H.; Poria, S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 9180–9192. [CrossRef]
18. Du, Z.; Wang, Y.; Yang, X.; Im, S.-K.; Wang, Z. MISA-GMC: An enhanced multimodal sentiment analysis framework with gated fusion and momentum contrastive modality relationship modeling. *Mathematics* **2026**, *14*, 115. [CrossRef]
19. Geng, X. Label distribution learning. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1734–1748. [CrossRef]
20. Wang, C.; Zhang, W.; Chen, G.; Dong, M.; He, T.; Jiang, X. Divide-and-conquer: Prompt-based distribution learning for multimodal sentiment analysis. *Inf. Fusion* **2026**, *133*, 104293. [CrossRef]
21. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.V.; Hinton, G.E.; Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
22. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.-P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [CrossRef]
23. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.-P. 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), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246. [CrossRef]
24. Liang, P.P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.Y.; Wu, P.; Lee, M.A.; Zhu, Y.; Salakhutdinov, R.; Morency, L.-P. MultiBench: Multiscale benchmarks for multimodal representation learning. In Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), Online, 6–14 December 2021.
25. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114. [CrossRef]
26. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.B.; Morency, L.-P. Efficient low-rank multimodal fusion with modality-specific factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2247–2256. [CrossRef]
27. Tsai, Y.-H.H.; Liang, P.P.; Zadeh, A.; Morency, L.-P.; Salakhutdinov, R. Learning factorized multimodal representations. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
28. Wu, T.; Peng, J.; Zhang, W.; Zhang, H.; Tan, S.; Yi, F.; Ma, C.; Huang, Y. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl. Based Syst.* **2022**, *235*, 107676. [CrossRef]
29. Sun, H.; Wang, H.; Liu, J.; Chen, Y.-W.; Lin, L. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 3722–3729. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.