Latent Distribution Decouple for Uncertain-Aware Multimodal Multi-label Emotion Recognition

Anonymous ACL submission

Abstract

Multimodal multi-label emotion recognition (MMER) aims to identify the concurrent presence of multiple emotions in multimodal data. Existing studies primarily focus on improving fusion strategies and modeling modality-to-label dependencies. However, they often overlook the impact of **aleatoric uncertainty**, which is the inherent noise in the multimodal data and hinders the effectiveness of modality fusion by introducing ambiguity into feature representations. To address this issue and effectively model aleatoric uncertainty, this paper proposes Latent emotional Distribution Decomposition with Uncertainty perception (LDDU) framework from a novel perspective of latent emotional space probabilistic modeling. Specifically, we introduce a contrastive disentangled distribution mechanism within the emotion space to model the multimodal data, allowing for the extraction of semantic features and uncertainty. Furthermore, we design an uncertainty-aware fusion multimodal method that accounts for the dispersed distribution of uncertainty and integrates distribution information. Experimental results show that LDDU achieves state-of-the-art performance on the CMU-MOSEI and M^{3}ED datasets, highlighting the importance of uncertainty modeling in MMER. We will release the related code.

1 Introduction

Human interactions convey multiple emotions through various channels: micro-expressions, vocal intonations, and text. Multimodal multi-label emotion recognition (MMER) seeks to identify multiple emotions (e.g., happiness, sadness) from multimodal data (e.g., audio, text, and video) (Zhang et al., 2021). It could support many downstream applications such as emotion analysis (Tsai et al., 2019), human-computer interaction (Chauhan et al., 2020), and dialogue systems (Ghosal et al., 2019).

The main topics of MMER lie in extracting emotion-relevant features by effectively fusing multimodal data and modeling modality-to-label dependencies (Zhang et al., 2021; Hazarika et al., 2020). To implement the fusion of multimodal data, some work (Zhang et al., 2020; Wang et al., 2024b) employed projection layers to mitigate the modality gap (Radford et al., 2021), while other methods (Zhang et al., 2021; Tsai et al., 2019) utilized attention mechanisms. Additionally, several studies (Hazarika et al., 2020; Zhang et al., 2022; Ge et al., 2023; Xu et al., 2024) decomposed modality features into common and private components. Recently, CARET (Peng et al., 2024) introduced emotion space modeling, where emotion-related features were extracted prior to fusion, achieving state-of-the-art performance. Regarding modality-to-label dependencies, many approaches (Zhang et al., 2021, 2022) leveraged Transformer decoders to capture the relationships between label semantic features and fused modality sequence features.

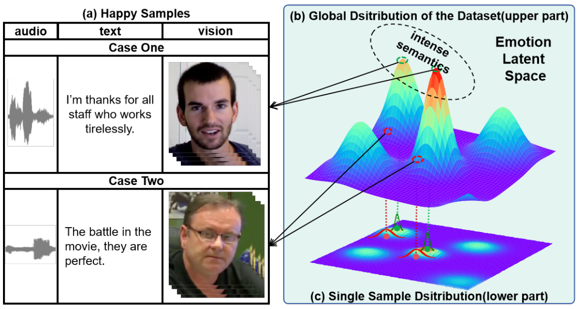


Figure 1: An illustration of aleatoric uncertainty in MMER task. When adopting Gaussian distribution modeling in latent emotion space, case two's semantic feature is more fuzzed with a larger variance due to cold speaking style than case one. Meanwhile, case one has stronger emotion intense located closer to the center of global distribution.

However, these approaches primarily focus on semantic features while overlooking **aleatoric uncertainty** (Kendall and Gal, 2017), which represents inherent noise in the data and is commonly modeled using multivariate Gaussian distributions

(Do, 2008) (for a detailed background, please refer to Appendix A.4). In the context of MMER, such uncertainty primarily arises from factors such as personalized expressions, variations in emotional intensity, and the blending of coexisting emotions (Zhao et al., 2021). For instance, as illustrated in Fig. 1, from a macroscopic perspective, both samples convey happiness, yet case one exhibits more pronounced facial expressions compared to case two. From a distributional perspective, case one demonstrates more concentrated semantic features near the center of the dataset's overall distribution, whereas case two presents features with greater variance, positioned farther from the center. This aleatoric uncertainty introduces ambiguity into semantic feature representations, thereby diminishing the effectiveness of modality fusion in existing MMER approaches (Gao et al., 2024).

To model aleatoric uncertainty in MMER, several challenges need to be addressed: (1) *How to represent aleatoric uncertainty:* Emotional cues are embedded in multimodal sequences, with each modality contributing differently to emotion expression, making it difficult to extract and disentangle emotional features. When modeled with multivariate Gaussian distributions, samples with the same label often cluster together despite semantic fuzziness. An effective distribution must capture both the central tendency and calibrate variance of emotional features, which is particularly challenging. (2) *How to integrate semantic features with aleatoric uncertainty:* Higher uncertainty leads to more dispersed distributions, complicating emotion recognition. Without calibrated uncertainty, semantic features can become ambiguous and less informative. Thus, effective strategies for calibrating and integrating uncertainty are crucial to ensure robust and discriminative emotion representations.

To address these challenges, we propose Latent Distribution Decouple for Uncertainty-Aware MMER (LDDU) from the perspective of latent emotional space probabilistic modeling. For the first challenge, to represent aleatoric uncertainty, LDDU extracts modality-related features using Q-Former-like alignment (Li et al., 2023). We then design a distribution decoupling mechanism based on Gaussian distributions to model uncertainty. To further enhance the distinguishability of these distributions, contrastive learning (Chen et al., 2020) is employed. For the second challenge, to effectively integrate the distributional information, we draw inspiration from uncertainty learning (Guo et al., 2017; Moon et al., 2020; Xu et al., 2024) and develop an uncertainty-aware fusion module, which is accompanied by uncertainty calibration. Experimental results on the CMU-MOSEI and M^{3}ED datasets show that LDDU achieves state-of-the-art performance. Specially, it surpasses strong baseline CARAT 4.3% miF1 on CMU-MOSEI under unaligned settings. In summary, the contributions of this work are as follows:

- We introduce latent emotional space probabilistic modeling for MMER. To the best of our knowledge, this is the first work to leverage emotion space distribution for capturing aleatoric uncertainty in MMER.
- We propose LDDU, which models the emotion space to extract emotion features, then uses contrastive disentagled learning to represent latent distributions and recognizes emotions by integrating both semantic features and calibrated uncertainty.
- Experiments on CMU-MOSEI and M^{3}ED datasets demonstrate that the proposed LDDU method achieves state-of-the-art performance, with mi-F1 improved 4.3% on the CMU-MOSEI unaligned data.

2 Related Work

Multimodal Multi-label Emotion Regression. It aims to infer human emotions from textual, audio, and visual sequences in video clips, often encompassing multiple affective states. The primary challenges in MMER is integrating multimodal data. Early studies like MISA (Hazarika et al., 2020) address modality heterogeneity by decoupled invariant and modality-specific features for fusion. MMS2S (Zhang et al., 2020) and HHMPN (Zhang et al., 2021) focused on modeling label-to-label and label-to-modality dependencies using Transformer and GNNs network. Recent approaches (Peng et al., 2024; Ge et al., 2023; Zhang et al., 2022) incorporated advanced training techniques; for example, TAILOR (Zhang et al., 2022) utilized adversarial learning to differentiate common and private modal features, while AMP (Wu et al., 2020) employed masking and parameter perturbation to mitigate modality bias and enhance robustness. However, these works all model from multimodal fusion instead of emotion latent space.

Uncertainty-aware Learning and Calibration.

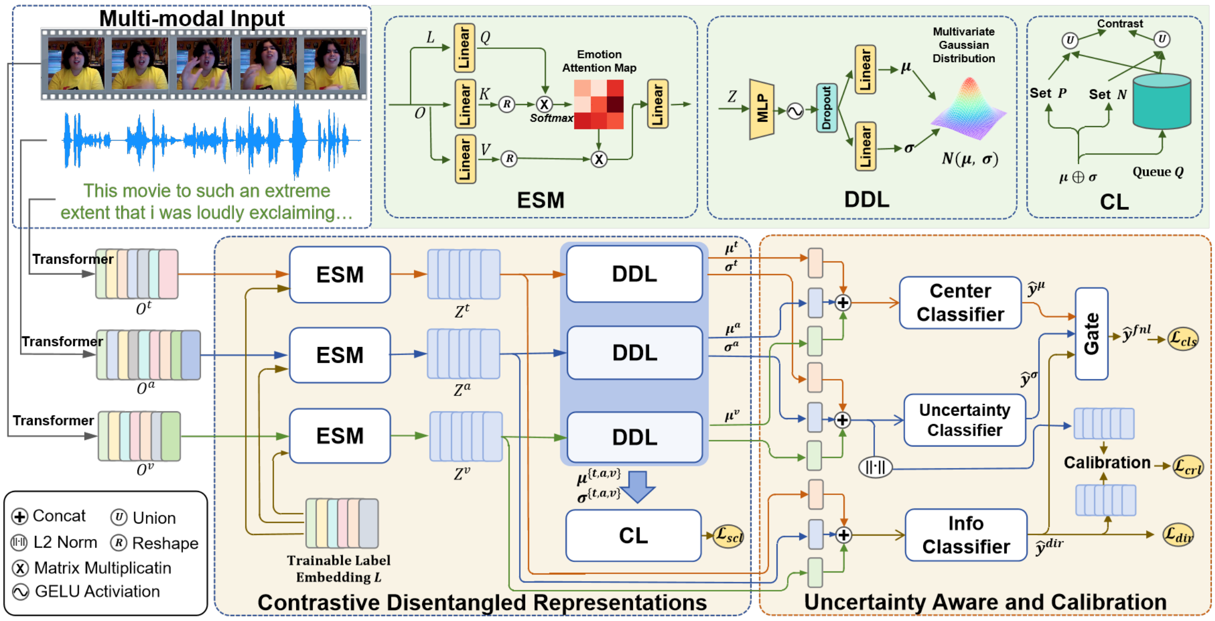


Figure 2: The proposed LDDU framework consists of three components: (1) the transformer-base unimodal extractor (2) a contrastive learning-based emotion space decomposition module , and (3) an uncertainty-aware fusion and uncertainty calibration module.

Deep models often overconfidently assign high confidence to incorrect predictions, making uncertainty-aware learning essential to ensure confidence accurately reflects prediction uncertainty (Guo et al., 2017). The primary goal is to calibrate model confidence to match the true probability of correctness. There are two main approaches: calibrated uncertainty (Guo et al., 2017) and ordinal or ranking-based uncertainty (Moon et al., 2020). Calibration methods, such as histogram binning, temperature scaling, and accuracy versus uncertainty calibration (Zadrozny and Elkan, 2001; Guo et al., 2017; Krishnan and Tickoo, 2020), align predicted confidence with actual correctness. Meanwhile, ranking-based methods like Confidence Ranking Loss (CRL) (Moon et al., 2020) enforce accurate confidence rankings among correctly classified samples based on feature distinctiveness.

Uncertainty-based Multimodal Fusion. Uncertainty learning enhances multimodal fusion across tasks. Subedar et al. (2019) employed Bayesian deep learning and AvU to guide fusion, while Xu et al. (2024) used temporal-invariant learning to reduce redundancy and noise, improving robustness. But these methods incorporate uncertainty without calibration. In contrast, COLD (Tellamekala et al., 2023) leveraged GURs to model feature distributions across modalities, quantifies modality contributions with variance norms, and integrated both calibrated and ranking-based uncertainty to regulate fusion variance. However, there hasn't exploration of uncertainty-aware for MMER.

3 Methodology

3.1 Preliminary

MMER is typically modeled as a multi-label task. Suppose $X^v \in \mathbb{R}^{s_v \times d_v}$, $X^a \in \mathbb{R}^{s_a \times d_a}$ and $X^t \in \mathbb{R}^{s_t \times d_t}$ denote the features of the text, visual, and audio modalities, respectively. In this context, s_v, s_a, s_t denote the length of the feature sequences, while d_v, d_a,d_t is the dimension of each features sequence. Given a multimodal sequential dataset in joint feature space $\mathcal{X}^{v,a,t}$, denoted as $\mathcal{D} = \{(X_i^v, X_i^a, X_i^t, y_i)\}_{i=1}^N$, the objective of the MMER is to learn the function $\mathcal{F}$: $\mathcal{D} \rightarrow \mathcal{Y}$. Here, N is the size of dataset $\mathcal{D}$ and X_i^v, X_i^a, X_i^t represent the visual, audio and textual sequences of the i-th sample. $\mathcal{Y} \in \mathbb{R}^q$ represent the emotion space containing q multiple coexisting emotion labels.

In this section, we describe LDDU framework, which comprises three components (in Figure 2).

3.2 Uni-Modal Feature Extractor

Follow the work of Peng et al. (2024), we conduct experiments on CMU-MOSEI (Zadeh et al., 2018b) and M3ED (Zhao et al., 2022) datasets. In these two benchmark, facial keypoints X^v via the MTCNN algorithm (Zhang et al., 2016), acoustic features X^a with Covarep (Degottex et al., 2014) and text features X^t of sample X are extracted using BERT (Yu et al., 2020). To capture content sequence dependencies, we employ n_v, n_a, and n_t Transformer layers as unimodal extractors, generating modality visual features $O^v \in \mathbb{R}^{s_v \times d_v}$, audio features $O^a \in \mathbb{R}^{s_a \times d_a}$, and text features $O^t \in \mathbb{R}^{s_t \times d_t}$. Each modality O^m is derived from its sequence data $[o_1^m, \ldots, o_{s_m}^m]$, $m \in \{v, a, t\}$.

3.3 Contrastive Disentangled Representation

3.3.1 Emotion Space Modeling

A primary challenge in emotion space modeling is the establishing emotion representations within a unified joint embedding space. Inspired by the Q-Former's structure (Li et al., 2023), we introduce trainable emotion embeddings $L = [l_1, l_2, ..., l_q]$, where each l_i represents an emotion and q is the number of label. Because emotion-related cues may be distributed across different segments of the sequential data, we employ an attention mechanism to automatically extract relevant features for each emotion. Since modality-related features O^m and L reside in different feature spaces, we use projection layers to compute the similarity a_{ij}^m between frame's feature o_j^m and the label l_i. After obtaining the similarity matrix $A^m = \{a_{ij}^m\}$, Y^m is projected to extract modality-specific label-related features $Z^m \in \mathbb{R}^{q \times d_h}$, where d_h is the dimension of modality-specific label-related features. This process could formalized as follows:

$$a_{ij}^m = \frac{exp(Proj(l_i)^T Proj(o_j^m))}{\sum_{j'=1}^{s_m} exp(Proj(l_i)^T Proj(o_{j'}^m))} \quad (1)$$

$$Z^m = Linear(A^m Proj(O^m)) \quad (2)$$

where $Proj$ represents the projection layer.

To facilitate the learning of emotion representations L, we concatenate the multimodal features of i-th sample into $F_{dir} = [Z_i^v, Z_i^a, Z_i^t]$ and process them with an MLP-based info classifier employing sigmoid activation functions to generate the final prediction $\hat{y}_i^{dir} = [\hat{y}_{i1}^{dir}, \ldots, \hat{y}_{iq}^{dir}]$. The loss function $\mathcal{L}_{dir}$ is defined as follows:

$$\mathcal{L}_{dir} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{BCE}(y_i, \hat{y}_i^{dir}) \quad (3)$$

where BCE(.) is the BCE loss.

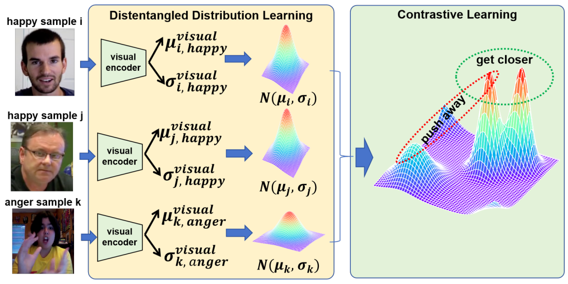


Figure 3: In the latent emotion space, we decouple emotion-related, modality-specific samples into separate distributions and use CL to group samples of the same category together while separating samples from different categories.

3.3.2 Contrastive Distentangled Distribution Modeling

This module is composed of distentangled distribution learning (DDL) and contrastive learning (CL). As illustrated in Figure 3, our architecture incorporates disentangled representation learning (DRL) (Wang et al., 2024b; Kingma, 2013) to establish label-related features into latent probabilistic distribution in emotion space. Specifically, we model multimodal emotion representations $[Z^v, Z^a, Z^t]$ as multivariate normal distributions $\mathcal{N}$. For each label-related features $[Z_i^v, Z_i^a, Z_i^t]$, we leverage an encoder (MLP in this paper) and two fully connected layers to obtain the latent distributions $\mathcal{N}(\mu_i^v, \sigma_i^v)$, $\mathcal{N}(\mu_i^a, \sigma_i^a)$ and $\mathcal{N}(\mu_i^t, \sigma_i^t)$. where μ_i^t represents the semantic features of the text modality for emotion label i (Tellamekala et al., 2023) and σ_i^t reflects the distribution region in latent space.

To ensure that the latent distribution $\mathcal{N}(\mu_i^m, \sigma_i^m)$ accurately captures feature differences for each label across modality m, we employ Contrastive Learning (CL). CL could groups similar samples together and enhances the model's ability to distinguish between different classes (He et al., 2020; Caron et al., 2020). Formally, given the variations in latent distributions across labels and modalities, we categorize them into $3q$ potential emotional distributions. For a batch of s_B samples $\mathcal{B}$, each sample generates $3q$ label-related and modality-specific emotion distributions, totaling $3 \times q \times s_B$ distributions. Each distribution in $\mathcal{B}^+$ is considered a positive sample if the related sample contains the corresponding emotion. For each positive distribution $e \in \mathcal{B}^+$, we identify its positive set $\mathcal{P}_e(\mathcal{B})$ and negative set $\mathcal{N}_e(\mathcal{B})$ based on labels.

Besides, we promote CL from the following two perspectives. First, Caron et al. (2020) observes that a larger batch size can enhance the network abilities by providing more diverse negative samples in CL. We introduce a queue Q of size s_q to store the most recent s_q emotion distributions. Thus, the final positive and negative sets for each emotion distribution become $\mathcal{P}_e(\mathcal{B} \cup \mathcal{Q})$ and $\mathcal{N}_e(\mathcal{B} \cup \mathcal{Q})$, respectively. Besides, similarity calculations between samples must consider both the centers and variances of the decoupled distributions. We represent the distribution e as follows:

$$e = \left(\frac{\mu_{e,1}}{|\mu_e|}, \ldots, \frac{\mu_{e,d_h/2}}{|\mu_e|}, \frac{\sigma_{e,1}}{|\sigma_e|}, \ldots, \frac{\sigma_{e,d_h/2}}{|\sigma_e|} \right) \quad (4)$$

Finally, we introduce the SupCon loss (Khosla et al., 2020) to for each emotion distribution:

$$\mathcal{L}_{scl}(e, \mathcal{B}^{+}) = \frac{-1}{|\mathcal{P}_e|} \sum_{e^{+} \in \mathcal{P}_e} \log \frac{e^{z(e,e^{+}/\tau)}}{\sum_{e' \in \mathcal{T}_e} e^{z(e,e')/\tau}} \quad (5)$$

where $\mathcal{T}_e = \mathcal{P}_e \cup \mathcal{N}_e$, and z is the similarity function between emotin distribution. To simplify the process, we calculated cosine similarity on the normalized distribution parameters:

$$z(e_1, e_2) = e_1^T e_2 \quad (6)$$

The final contrastive loss for the entire batch is:

$$\mathcal{L}_{scl} = \sum_{e \in \mathcal{B}^{+}} \mathcal{L}_{scl}(e, \mathcal{B}^{+}) \quad (7)$$

3.4 Uncertainty Aware and Calibration

3.4.1 Uncertainty-Awared Multimodal Fusion

After modeling the emotional space, it's crucial to integrate latent semantic features with the distribution uncertainty information. We use variance to represent the distribution uncertainty in latent space, as it reflects the degree of dispersion and distribution region. Meanwhile, the center feature represents the semantic features of a sample (Gao et al., 2024; Tellamekala et al., 2023; Xu et al., 2024). We hypothesize that when a sample has high aleatoric uncertainty, its semantic features become fuzzier, and the distribution region in latent space becomes more discriminative for emotion recognition. Conversely, when aleatoric uncertainty is low, the semantic features are more discriminative, and the distribution region is narrower. Therefore, the fusion of center feature and variance should depend on the level of aleatoric uncertainty.

Firstly, we introduce the i-th sample's prediction $\hat{y}_i^{dir}$ of Info Classifier to quantify uncertainty. Kendall and Gal (2017) pointed out that aleatoric uncertainty can be measured by the prediction difficulty of the sample. Specifically, if Z_i correctly classified by Info Classifier while Z_j is misclassified and needs to be decoupled for further classification. We infer that the j-th sample exhibits higher aleatoric uncertainty(i.e., is less informative). Consequently, the uncertainty can be represented as $d(\hat{y_i}^{dir}, y_i)$, where $\hat{y_i}^{dir}$ is the prediction of Z_i.

Then, we integrate the distribution's information by fusing multimodal data. After decoupled, the samples are represented as latent distributions $\mathcal{N}\left(E^{v,a,t}, M^{v,a,t}\right)$ where E^m=$[\mu_1^m, ..., \mu_q^m]$ and M^m=$[\sigma_1^m, ..., \sigma_q^m]$ for each modality m. Since $E^{v,a,t}$ and $M^{v,a,t}$ have different semantics, we implement late fusion using gate network. Operationally, (E^v,E^a,E^t) and (M^v,M^a,M^t) are concatenated and passed through final classifier to obtain the predictions $\hat{y_i}^{\mu}$ and $\hat{y_i}^{\sigma}$. Semantic mean vector and the variance are dynamically fused according to uncertainty score:

$$\hat{y}_i^{fnl} = d(\hat{y}_i^{dir}, y_i)\hat{y}_i^{\mu} + (1 - d(\hat{y}_i^{dir}, y_i))\hat{y}_i^{\sigma} \quad (8)$$

For a batch of data with size s_B, the loss function is as follows:

$$\mathcal{L}_{cls} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathrm{BCE}(\hat{y}_i^{fnl}, y_i) \quad (9)$$

3.4.2 Uncertainty Calibration

In this section, we impose ordinality constraint (Moon et al., 2020) to model the relationship between uncertainty and distribution variance. When well-calibrated, the uncertainty score acts as a proxy for the correctness likelihood of its prediction for the latent distribution. In other words, well-calibrated uncertainty indicates the expected estimation error, i.e., how far the predicted emotion is expected to lie from its ground truth.

It has been confirmed that: frequently forgotten samples are harder to classify, while easier samples are learned earlier in training (Toneva et al., 2018; Geifman et al., 2018). As a result, to represent the correctness likelihood values, we use the proportion of samples r_i correctly predicted by the Info Classifier during the SGD process (Shamir and Zhang, 2013; Xu et al., 2024).

In our approach, the variance $\sigma_i = (\sigma_i^v, \sigma_i^a, \sigma_i^t)$ and the prediction error $d(\hat{y}_i^{dir}, y_i)$ from the Info Classifier are strongly correlated with the correctness likelihood values of emotion classification.

Thus, the calibration can be formulated as follows:

$$\arg\max \ \mathrm{Corr}(rk(\tfrac{1}{\|\sigma_i\|_2},\tfrac{1}{\|\sigma_j\|_2}),rk(r_i,r_j)) \tag{10}$$

$$\arg\max \ \mathrm{Corr}(rk(1-d_i,1-d_j),rk(r_i,r_j)) \tag{11}$$

where rk is ranking and $Corr$ is correlation. When the sample contain high uncertainty, the latent distribution variance σ_i and the prediction error $d_i = d(\hat{y}_i^{dir}, y_i)$ tend to be large, while r_i tend to be small. Conversely, when the uncertainty is small, these features are reversed.

For a batch of size s_B, we we compute the variance norm S, distance vector D, and proxy vector R for each sample:

$$S = [\frac{1}{||\sigma_1||_2}, \frac{1}{||\sigma_2||_2}, .., \frac{1}{||\sigma_{s_B}||_2}] \tag{12}$$

$$D = [1-d_1, 1-d_2, .., 1-d_{s_B}] \tag{13}$$

$$R = [r_1, r_2, .., r_{s_B}] \tag{14}$$

In order to establish the ranking constraints among S, D and R, we impose ordinality constraints based on soft-ranking (Tellamekala et al., 2023; Bruch et al., 2019). Our method employs bidirectional KL divergence to assess mismatching between the softmax distributions of pairs (S, R) and (D, R). Consequently, ordinality calibration loss $\mathcal{L}_{ocl}$ can be calculated as follows:

$$\begin{aligned}\mathcal{L}_{ocl} = &KL(P_D||P_R) + KL(P_R||P_D) \\ &+ KL(P_S||P_R) + KL(P_R||P_S)\end{aligned} \tag{15}$$

where P_D, P_R, and P_S represent the softmax distributions of features S, R, and D, respectively.

Overall, in the whole training process, the training loss of LDDU is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{ocl} + \beta\mathcal{L}_{scl} + \gamma\mathcal{L}_{dir} \tag{16}$$

where λ, β, and γ are hyperparameters controlling the weight of each regularization constraint.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We validate the proposed method LDDU on two benchmark: CMU-MOSEI (Zadeh et al., 2018b) and M3ED (Zhao et al., 2022). CMU-MOSEI consists of 23,453 video segments across 250 topics. Each segment is labeled with multiple emotions, including happiness, sadness, anger, disgust, surprise, and fear. The M^{3}ED dataset, designed for dialog emotion recognition, offers greater volume and diversity compared to IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2018). It includes 24,449 segments, capturing diverse emotional interactions with seven emotion categories: the above six emotions with neutral. Following previous work (Zhang et al., 2021, 2022; Wu et al., 2020; Peng et al., 2024), in the experiments, we evaluate model performance using accuracy (Acc), precision (P), recall (R), and micro-F1 score (miF1).

Baselines. We compare the LDDU model with two types methods: traditional multimodal methods and multimodal large language model (MLLM) methods. Traditional methods include DFG (Zadeh et al., 2018a), RAVEN (Wang et al., 2019), MulT (Tsai et al., 2018), MISA (Hazarika et al., 2020), MMS2S (Zhang et al., 2020), HHMPN (Zhang et al., 2021), TAILOR (Zhang et al., 2022), AMP (Wu et al., 2020), and CARAT (Peng et al., 2024).

Furthermore, given the significant success of MLLMs in multimodal tasks, we compare LDDU with MLLMs such as GPT-4o (*gpt-4o-2024-11-20*) (Achiam et al., 2023), Qwen2-VL-7B (Wang et al., 2024a), and AnyGPT (Zhan et al., 2024). They respectively correspond to the open-source paradigm, closed-source paradigm, and the omni large language model (LLM). We conduct experiments using raw video clips (treated as unalgined data) from the CMU-MOSEI dataset, maintaining consistent prompts and experimental settings with the framework proposed by Lian et al. (2024). Details of the prompts are provided in Appendix A.3.

In addition, we conducted a comprehensive comparison between the LDDU and existing multi-label classification (MLC) approaches including both classical methods: BR (Boutell et al., 2004), LP (Tsoumakas and Katakis, 2008), CC (Read et al., 2011); and single-modality methods: SGM (Yang et al., 2018), LSAN (Xiao et al., 2019), ML-GCN (Wu et al., 2019), please see Appendix A.2 and Table 4 for full comparisons.

Implementation Details. We set $\lambda = 0.1$, $\beta = 0.8$, and $\gamma = 0.1$, with a batch size of 128. The learning rate is 2e-5 with 30 epochs. More details of all experiences are shown in Appendix A.1.

4.2 Experimental Results

Main Results. In Table 1 and Table 2, we compare the performance of our method with various baseline approaches on the CMU-MOSEI and M^{3}ED datasets. Different from most baseline methods listed in Table 1, which use the CTC (Graves

Table 1: Performance comparison on the CMU-MOSEI dataset under aligned and unaligned settings. As LLM-based methods process raw video segments, aligned results are unavailable. Best results are red, second-best are blue. A full comparison between multimodal methods, MLLMs and classical methods is in Appendix A.2.

Approaches	Methods	Aligned				Unaligned			
		Acc	P	R	miF1	Acc	P	R	miF1
LLM-based	GPT-4o	—	—	—	—	0.352	0.583	0.252	0.196
	Qwen2-VL-7B	—	—	—	—	0.422	0.520	0.355	0.355
	AnyGPT	—	—	—	—	0.134	0.251	0.445	0.321
Multimodal	DFG	0.396	0.595	0.457	0.517	0.386	0.534	0.456	0.494
	RAVEN	0.416	0.588	0.461	0.517	0.403	0.633	0.429	0.511
	MulT	0.445	0.619	0.465	0.501	0.423	0.636	0.445	0.523
	MISA	0.430	0.453	**0.582**	0.509	0.398	0.371	**0.571**	0.450
	MMS2S	0.475	0.629	0.504	0.516	0.447	0.619	0.462	0.529
	HHMPN	0.459	0.602	0.496	0.556	0.434	0.591	0.476	0.528
	TAILOR	0.488	0.641	0.512	0.569	0.460	0.639	0.452	0.529
	AMP	0.484	0.643	0.511	0.569	0.462	**0.642**	0.459	0.535
	CARAT	**0.494**	**0.661**	0.518	**0.581**	**0.466**	**0.652**	0.466	**0.544**
	LDDU	**0.494**	**0.647**	**0.531**	**0.587**	**0.496**	0.638	**0.543**	**0.587**

Table 2: Performance comparison on the M^3ED dataset.

Methods	Acc	P	R	miF1
MMS2S	0.645	0.813	0.737	0.773
HHMPN	0.648	0.816	0.743	0.778
TAILOR	0.647	0.814	0.739	0.775
AMP	0.654	0.819	0.748	0.782
CARAT	**0.664**	**0.824**	**0.755**	**0.788**
LDDU	**0.690**	**0.843**	**0.774**	**0.807**

Table 3: Ablation tests on the aligned CMU-MOSEI.

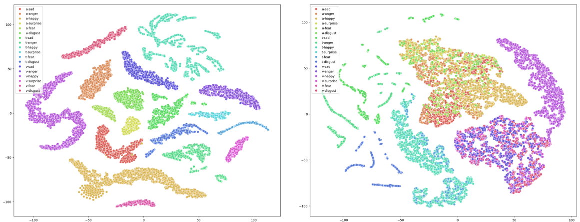
Approach	Acc	P	R	miF1
(1) w/o ESM	0.478	0.663	0.510	0.577
(2) w/o $\mathcal{L}_{dir}$	0.491	0.656	0.521	0.580
(3) w/o $\mathcal{L}_{scl}$	0.483	**0.679**	0.498	0.575
(4) w/o queue $\mathcal{Q}$	0.487	0.655	0.487	0.578
(5) w/o variance μ	0.483	0.628	0.536	0.578
(6) w/o center σ	0.492	0.647	0.527	0.581
(7) w/o $\mathcal{L}_{ocl}$	0.483	0.672	0.510	0.580
(8) ow Corr(S, R)	0.484	0.641	0.532	0.581
(9) ow Corr(D, R)	0.490	0.647	0.533	0.584
(10) ow Corr(D, S)	0.492	0.633	0.538	0.582
(11) $\mathcal{L}_{cls}$ w/o $\hat{y}^{\mu}$	0.485	0.666	0.510	0.578
(12) $\mathcal{L}_{cls}$ w/o $\hat{y}^{\sigma}$	0.494	0.622	**0.543**	0.580
(12) LDDU	**0.494**	0.647	0.531	**0.587**

et al., 2006) module to align non-aligned datasets, LDDU performs greatly better on unaligned data without relying on the CTC module. The emotion extraction network in LDDU directly extracts modality-specific features related to the labels from the sequence data, unaffected by the varying sequence lengths across modalities.

Based on Tables 1 and 2, we can draw the following observations: (1) LDDU outperforms other baseline methods on more crucial metrics such as mi-F1 and accuracy(acc) although recall and precision scores are not the highest on the CMU-MOSEI dataset. Notably, LDDU achieved balanced performance on both aligned and unaligned datasets, with unaligned's accuracy improved by 3% and unaligned's mi-F1 increased by 4.3%. This demonstrates that by modeling the emotional space rather than sequence features, LDDU can better capture emotion-related features. (2) LDDU also achieved significant improvements across all metrics on the M^3ED dataset, confirming the robustness of our model. (3) TAILOR, CAFET, and the proposed LDDU approach performed better by separating features, emphasizing the importance of considering each modality's unique contributions to emotion recognition in MMER tasks. (4) While MLLMs are excellent at video understanding, the proposed method significantly outperforms MLLMs. This maybe because their ability to capture finer-grained emotional information is limited and smaller models outperform them in this regard.

Ablation Study. To elucidate the significance of each component of proposed methods, we compared LDDU against various ablation variants. As shown in Table 3, where "w/o" means removing, "ow" denotes only existing, "w/o ESM" denotes removing trainable feature L. "w/o σ, μ" respectively means only consider variances or centers during CL, "w/o Corr(S,R), Corr(D,R)" denotes ignoring the calibration of (S,R) or (D,R), "$\mathcal{L}_{cls}$ w/o $\hat{y}^{\sigma}$, $\hat{y}^{\mu}$" means final classification without variance or semantic center features. We could find:

1) *Effect of emotion space modeling (ESM).* We replaced ESM with MLP-based attention in (1) and dropped the loss $\mathcal{L}_{dir}$ in (2). Both of them illustrated the trainable features L with supervisory



(a). t-SNE with CL (b). t-SNE without CL

Figure 4: The t-SNE visualization of embedding with /without CL datasets. Different colors represent different label-related modality-special features of samples.

signals from $\mathcal{L}_{dir}$ can learn more distinguishable features of raw multimodel sequences

2) *Effect of contrastive learning.* We compared LDDU with the variants without $\mathcal{L}_{scl}$ in (3). Performance degradation across metrics confirms the essential role of CL in decoupling. (4) is better than (3), which illustrates a larger batch size can enhance CL. Further, (5) and (6) demonstrates when computing similarity between distributions, both mean value and variance should be considered.

3) *Effect of uncertainty calibration.* Compared with variants without cilbration, the implementations of constraints (8, 9, 10, 12) show enhanced performance. This calibration aligned the variance with uncertainty, generating better predictions.

4) *Effect of uncertainty-awared fusion.* To modeling aleatoric uncertainty, we integrated the semantic features with the distribution's regional information. (11) and (12) illustrates that both of them contributes to the final classification.

4.3 Further Analysis

Visualization of Emotion Distribution. To evaluate the effectiveness of Contrastive Learning (CL), we used t-SNE (van der Maaten and Hinton, 2008) to visualize latent emotion distributions from the CMU-MOSEI test set, excluding samples without specific emotions. As shown in Fig. 4, panels (a) and (b) display distributions with and without CL, respectively. Without CL, a clear modality gap exists and intra-modality distributions lack distinctiveness. With CL, the $3 \times nl$ emotion distributions across labels and modalities are distinctly separated, enhancing their distinguishability. Consequently, LDDU leveraging CL can more effectively learn emotion distributions across modalities within the joint emotional space, with each cluster representing a specific emotion.

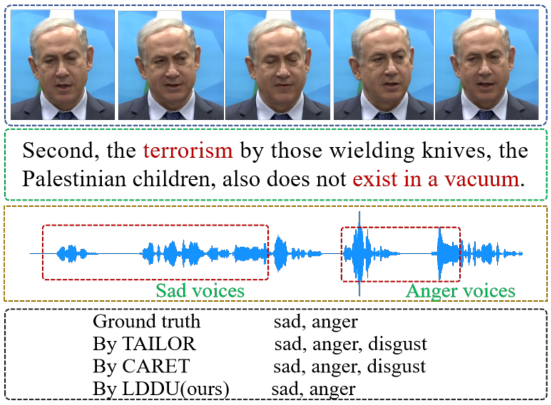


Figure 5: The case of emotion recognition by multiple methods. Visual and acoustic modalities revealed a shift from sadness to anger, while the textual modality explicitly indicated anger-related expressions.

Case Study. To validate LDDU's effectiveness, Figure 5 illustrates a representative case where visual/acoustic modalities indicate a transition from sadness to anger, while textual modality explicitly signals anger. While all methods accurately detected sadness and anger, TAILOR and CARET falsely predicted disgust due to its ambiguous emotional cues in overlapping scenarios. In contrast, LDDU effectively modeled emotion-specific uncertainties through latent space Gaussian distributions (distance vector D's value: sad: 0.23, anger: 0.41, disgust: 0.82). We further computed emotion correlation matrices (M_1–M_3 for methods, M_0 for ground truth) and measure their cosine similarities with M_0: LDDU achieved 96.7% (vs. 93.3% for TAILOR, 96.1% for CARET), demonstrating superior capability in capturing inter-emotion dependencies. More cases are shown in Appendix A.5.

5 Conclusion

We propose LDDU, a framework that captures aleatoric uncertainty in MMER through latent emotional space probabilistic modeling. By disentangling semantic features and uncertainty using Gaussian distributions, LDDU mitigates ambiguity arising from variations in emotional intensity and overlapping emotions. Furthermore, an uncertainty-aware fusion module adaptively integrates multimodal features based on their distributional uncertainty. Experimental results on CMU-MOSEI and M^3ED demonstrate that LDDU achieves state-of-the-art performance. This work pioneers probabilistic emotion space modeling, providing valuable insights into uncertainty-aware affective computing.

Limitation

While LDDU demonstrates promising performance in MMER, several problems remain to discuss. LDDU models emotion uncertainty using Gaussian distributions in the latent emotion space, effectively capturing inherent ambiguity. However, it does not explicitly utilize emotion intensity labels, as provided in the CMU-MOSEI dataset (quantized into 0, 0.3, 0.6, and 1.0 levels). While this omission ensures fair comparisons with prior work (e.g., TAILOR, CARET), it also limits LDDU's ability to precisely distinguish emotions of varying intensities. As a result, the model may be less effective in disambiguating overlapping emotions, particularly in tasks requiring fine-grained intensity differentiation. Integrating explicit intensity supervision in future iterations could further refine LDDU's predictive capability.

Ethical Considerations

Ethical considerations are crucial in multimodal emotion recognition research, particularly with sensitive human data like emotional expressions. In our study, we ensure that all datasets, including CMU-MOSEI and M^{3}ED, are publicly available and anonymized to protect individuals' privacy.

While our method advances emotion recognition in areas such as human-computer interaction, we acknowledge the potential for misuse, such as manipulation or surveillance. We emphasize the responsible use of these technologies, ensuring they are deployed in contexts that respect privacy.

Additionally, emotional expressions vary across cultures and individuals, and our model may not fully capture this diversity. We recommend expanding datasets to include a wider range of cultural contexts to avoid biases and misinterpretations.

Finally, we commit to transparency by making our code publicly available for further scrutiny and improvement, ensuring our research aligns with ethical principles and benefits society.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. 2004. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771.

Sebastian Bruch, Xuanhui Wang, Michael Bendersky, and Marc Najork. 2019. An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pages 75–78.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4351–4360.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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

Chuong B Do. 2008. The multivariate gaussian distribution. *Section Notes, Lecture on Machine Learning, CS*, 229.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26876–26885.

Shiping Ge, Zhiwei Jiang, Zifeng Cheng, Cong Wang, Yafeng Yin, and Qing Gu. 2023. Learning robust multi-modal representation for multi-label emotion recognition via adversarial masking and perturbation. In *Proceedings of the ACM Web Conference 2023*, pages 1510–1518.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. 2018. Bias-reduced uncertainty estimation for deep neural classifiers. *arXiv preprint arXiv:1805.08206*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranganath Krishnan and Omesh Tickoo. 2020. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33:18237–18248.

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

Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. 2024. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. 2020. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pages 7034–7044. PMLR.

Cheng Peng, Ke Chen, Lidan Shou, and Gang Chen. 2024. Carat: Contrastive feature reconstruction and aggregation for multi-modal multi-label emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14581–14589.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85:333–359.

Ohad Shamir and Tong Zhang. 2013. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR.

Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. 2019. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6301–6310.

Mani Kumar Tellamekala, Shahin Amiriparian, Björn W Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. 2023. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Grigorios Tsoumakas and Ioannis Katakis. 2008. Multi-label classification: An overview. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, pages 64–74.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xin Wang, Hong Chen, Zihao Wu, Wenwu Zhu, et al. 2024b. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7216–7223.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969.

Xuan Wu, Qing-Guo Chen, Yao Hu, Dengbao Wang, Xiaodong Chang, Xiaobo Wang, and Min-Ling Zhang. 2019. Multi-view multi-label learning with view-specific information extraction. In *IJCAI*, pages 3884–3890.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 466–475.

Guoyang Xu, Junqi Xue, Yuxin Liu, Zirui Wang, Min Zhang, Zhenxi Song, and Zhiguo Zhang. 2024. Semantic-guided multimodal sentiment decoding with adversarial temporal-invariant learning. *arXiv preprint arXiv:2409.00143*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: sequence generation model for multi-label classification. *arXiv preprint arXiv:1806.04822*.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018a. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *Preprint*, arXiv:2402.12226.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3584–3593.

Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14338–14346.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9100–9108.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint arXiv:2205.10237*.

Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. 2021. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6):59–73.

A Appendix

A.1 Implementation Details

We set $\lambda = 0.1$, $\beta = 0.8$, and $\gamma = 0.1$, with a batch size of 128. For the uni-modal extraction network, each Transformer consists of 3 layers ($l_a = l_v = l_t = 3$). The hidden dimensions are 256 for feature Y and 128 for feature Z. The latent emotion distribution has a dimension of 64 for both the distribution centers and variance vectors. The contrastive learning queue $\mathcal{Q}$ is sized at 8192. The number of labels (q) is 6 for CMU-MOSEI and 7 for M^{3}ED. We optimize all model parameters using the Adam optimizer (Kingma, 2014) with a learning rate of 2×10^{-5} and a cosine decay schedule with a warm-up rate of 0.1. All experiments are conducted on a single GTX A6000 GPU using grid search.

A.2 More Compared Baselines

Despite the advancements in LLM-based and multimodal methods, we conducted a comprehensive and comparative analysis between the LDDU model and existing multi-label classification (MLC) approaches. This comparison includes both classical methods (BR (Boutell et al., 2004), LP (Tsoumakas and Katakis, 2008), CC (Read et al., 2011)) and single-modality methods (SGM (Yang et al., 2018), LSAN (Xiao et al., 2019), ML-GCN (Wu et al., 2019)). The experimental results are presented in Table 4.

A.3 Prompts of MLLM

In this study, we evaluated three multimodal models (GPT-4o, Qwen2-VL-7B, and AnyGPT), using video clips with an average duration of 7–8 seconds. GPT-4o and Qwen2-VL-7B exhibit strong visual understanding capabilities, representing closed-source and open-source multimodal large language models (MLLMs), respectively. AnyGPT is a versatile any-to-any MLLM capable of processing images, text, and audio. Since all these MLLMs adopt end-to-end architectures, we ensured computational efficiency and consistency by uniformly sampling 8 frames per video clip as input for inference. The specific prompts designed for each model, including task descriptions and format requirements, are detailed in Figure 6.

A.4 Detailed Info of Uncertainty Caliration

To enhance readers' understanding of aleatoric uncertainty and uncertainty correction, we provide additional supplementary materials.

A.4.1 Aleatoric Uncertainty in MMER

Aleatoric uncertainty refers to the inherent variability or noise present in the data, arising from factors beyond the model's control. In the context of emotion recognition, it stems from factors such as variations in emotional intensity, individual differences, and the blending of multiple emotions. This form of uncertainty is intrinsic to the data itself.

In multimodal emotion recognition (MMER), aleatoric uncertainty becomes particularly evident when the same emotion is expressed by different individuals. For example, a person may express happiness through a broad smile (visual modality) but with a neutral tone of voice (audio modality), reflecting differences in emotional intensity and expression. These inconsistencies can introduce conflicting cues that complicate the emotion recognition process. Furthermore, datasets like CMU-MOSEI also contain varying levels of emotion intensity, further contributing to aleatoric uncertainty.

This type of uncertainty is not confined to emotion recognition alone. In computer vision (CV), it can manifest as blurry faces or imprecise object localization, introducing uncertainty in tasks like object detection. In natural language processing (NLP), aleatoric uncertainty arises from ambiguities in language, where word meanings can shift based on contextual factors. In all these scenarios, probabilistic models are employed to capture and account for such inherent uncertainty, thereby enhancing the robustness of systems in diverse, real-world environments.

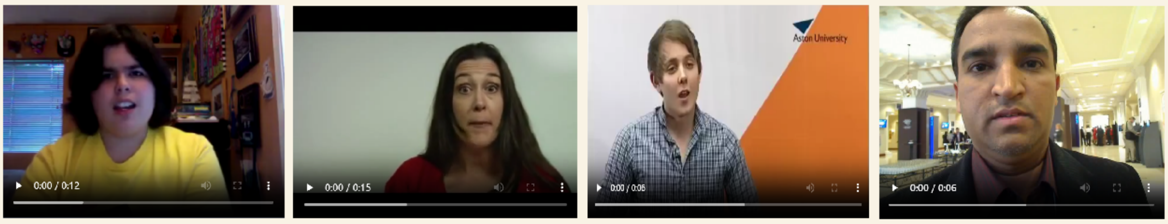
A.4.2 Uncertainty Calibration

Uncertainty Calibration. Uncertainty Calibration refers to the process of adjusting model predictions to more accurately reflect the true uncertainty associated with them. In machine learning and deep learning, models often provide predictions accompanied by an associated uncertainty; however, these predictions are not always well-calibrated. In other words, the model may exhibit excessive confidence in certain predictions, even when the true uncertainty is high, or it may fail to properly estimate its own uncertainty.

The primary objective of uncertainty calibration is to align the predicted uncertainty with the actual likelihood of a prediction being correct. In practical terms, this means that if a model is 90%

Table 4: Performance comparison on the CMU-MOSEI dataset under aligned and unaligned settings. As LLM-based methods process raw video segments, aligned results are unavailable. Best results are red, second-best are blue.

Approaches	Methods	Aligned				Unaligned			
		Acc	P	R	miF1	Acc	P	R	miF1
LLM-based	GPT-4o	—-	—-	—-	—-	0.352	0.583	0.252	0.196
	Qwen2-VL-7B	—-	—-	—-	—-	0.422	0.520	0.355	0.355
	AnyGPT	—-	—-	—-	—-	0.134	0.251	0.445	0.321
Classical	BR	0.222	0.309	0.515	0.386	0.233	0.321	0.545	0.404
	LP	0.159	0.231	0.377	0.286	0.185	0.252	0.427	0.317
	CC	0.225	0.306	0.523	0.386	0.235	0.320	0.550	0.404
Deep-based	SGM	0.455	0.595	0.467	0.523	0.449	0.584	0.476	0.524
	LSAN	0.393	0.550	0.459	0.501	0.403	0.582	0.460	0.514
	ML-GCN	0.411	0.546	0.476	0.509	0.437	0.573	0.482	0.524
Multimodal	DFG	0.396	0.595	0.457	0.517	0.386	0.534	0.456	0.494
	RAVEN	0.416	0.588	0.461	0.517	0.403	0.633	0.429	0.511
	MulT	0.445	0.619	0.465	0.501	0.423	0.636	0.445	0.523
	MISA	0.430	0.453	**0.582**	0.509	0.398	0.371	**0.571**	0.450
	MMS2S	0.475	0.629	0.504	0.516	0.447	0.619	0.462	0.529
	HHMPN	0.459	0.602	0.496	0.556	0.434	0.591	0.476	0.528
	TAILOR	0.488	0.641	0.512	0.569	0.460	0.639	0.452	0.529
	AMP	0.484	0.643	0.511	0.569	0.462	**0.642**	0.459	0.535
	CARAT	**0.494**	**0.661**	0.518	**0.581**	**0.466**	**0.652**	0.466	**0.544**
	LDDU	**0.494**	**0.647**	**0.531**	**0.587**	**0.496**	0.638	**0.543**	**0.587**



Prompt: Please play the role of a video expression classification expert. We provide 4 videos, each video with the speaker's spoken words and 8 temporally uniformly sampled frames. For each video and its corresponding text, please judge whether provided emotion categories are in the video based on the spoken words and corresponding frames and give out the existing categories. Please note that each video may contain multiple emotions.

Here are the optional categories: [“happy”, “sad”, “anger”, “surprise”, “disgust”, “fear”].

Please ignore the speaker's identity and focus on their emotions both in the sampled frames and in speech context and ignore the speaker's identity and focus on their emotions in the sampled frames and spoken words.

The output format should be {{'name':, 'result':}} for these {len(video_paths)} videos.
For example:

Video x:
- Content: "And I give this movie a two out of five because it's a bad movie, that's no surprise."
- Emotions in frames: emotion1, emotion2
- Emotions in spoken words: emotion1, emotion3
- Recognized categories: `{'name': 'Video x', 'result': ['emotion1', 'emotion2', 'emotion3']}`

Figure 6: Prompts of MLLMs.

confident in its prediction, it should be correct approximately 90% of the time over a large number of predictions. This calibration process is particularly critical in domains such as emotion recognition, medical diagnosis, and autonomous driving, where accurate uncertainty estimates are essential for reliable decision-making. Several methods can be employed for uncertainty calibration, including temperature scaling, Platt scaling, and Bayesian approaches.

Ordinality Constraint. Ordinality Constraint refers to a form of uncertainty calibration that is based on the ranking of classes. This method assumes that the relationship between classes or labels follows a natural ordinal structure, where labels have an inherent order. For instance, in sentiment analysis, labels such as "very negative," "negative," "neutral," "positive," and "very positive" exhibit a natural progression from negative to positive sentiment. Ordinality constraints ensure that the model's predicted probabilities reflect this ranking, adjusting the output so that predictions align with the ordered nature of the classes.

In our proposed approach, the ordinality constraint is applied to rank the uncertainty of predictions across different labels. By incorporating this constraint, we ensure that the model not only outputs probabilities but also ranks the classes in a manner that respects their inherent order.

A.4.3 Uncertainty Caliration in LDDU

Since networks learning variance and mean vectors share similar structures, variance and mean tend to converge and surface feature space collapse without constraints. The key is to ensure that variance vectors accurately reflect uncertainty level. We introduce an ordinality (ranking) constraint (Moon et al., 2020) to solve this problem. As shown in Equation 1, ordinality constraint requires predicted confidence κ should correspond to the probability $\mathcal{P}$ of correct prediction. In our approach, the variance $\sigma_i = (\sigma_i^v, \sigma_i^a, \sigma_i^t)$ and the prediction error $d(\hat{y}_i^{dir}, y_i)$ from the Info Classifier jointly represent the sample's confidence. The main challenge is establishing reliable proxy features for $\mathcal{P}$. Inspired by CRL (Xu et al., 2024), we use the proportion of samples r_i correctly predicted by the Info Classifier during the SGD (Shamir and Zhang, 2013) process as a proxy for $\mathcal{P}$. Empirical findings from Toneva et al. (2018) and Geifman et al. (2018) support our hypothesis: frequently forgotten samples are harder to classify, while easier samples are learned earlier in training.

When the sample contain high uncertainty, the latent distribution variance σ_i and the prediction error $d_i = d(\hat{y}_i^{dir}, y_i)$ tend to be large, while r_i tend to be small. Conversely, when the uncertainty is small, these features are reversed. Therefore, the ordinality constraint is:

$$max\ Corr(rk(\frac{1}{||\sigma_i||_2}, \frac{1}{||\sigma_j||_2}),\ rk(r_i, r_j)) \tag{17}$$

$$argmax\ Corr(rk(1-d_i, 1-d_j),\ rk(r_i, r_j)) \tag{18}$$

where $Corr$ represents correlation and rk demotes ranking. In this paper, we impose ordinality constraints based on soft-ranking(Tellamekala et al., 2023; Bruch et al., 2019). While (Tellamekala et al., 2023) uses KL divergence to measure mismatching of softmax distributions and (Bruch et al., 2019) applies softmax cross-entropy for ordinal regression, our method employs bidirectional KL divergence to assess mismatching between the softmax distributions

For a batch of size s_B, we compute the variance norm S, distance vector D, and proxy vector R for each sample:

$$S = [\frac{1}{||\sigma_1||_2}, \frac{1}{||\sigma_2||_2}, .., \frac{1}{||\sigma_{s_B}||_2}] \tag{19}$$

$$D = [1-d_1, 1-d_2, .., 1-d_{s_B}] \tag{20}$$

$$R = [r_1, r_2, .., r_{s_B}] \tag{21}$$

Inspired by (Tellamekala et al., 2023; Bruch et al., 2019), we impose ordinality constraints based on soft-ranking. While (Tellamekala et al., 2023) uses KL divergence to measure mismatching of softmax distributions and (Bruch et al., 2019) applies softmax cross-entropy for ordinal regression, our method employs bidirectional KL divergence to assess mismatching between the softmax distributions of pairs (S, R) and (D, R). Consequently, ordinality calibration loss $\mathcal{L}_{ocl}$can be calculated as follows:

$$\begin{aligned}\mathcal{L}_{ocl} =& KL(P_D||P_R) + KL(P_R||P_D) \\ &+ KL(P_S||P_R) + KL(P_R||P_S)\end{aligned} \tag{22}$$

where P_D, P_R, and P_S represent the softmax distributions of features S, R, and D, respectively.

In summary, the total training loss is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{ocl} + \beta\mathcal{L}_{scl} + \gamma\mathcal{L}_{dir} \tag{23}$$

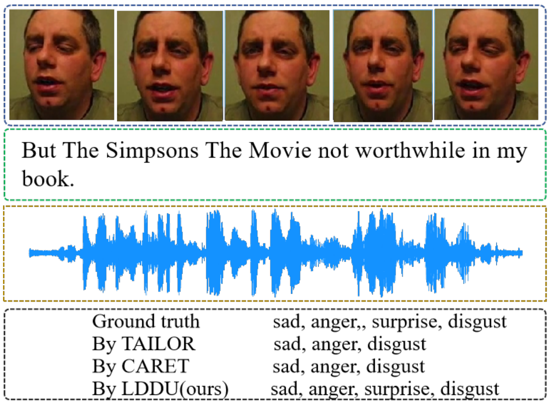


Figure 7: The case of emotion recognition by multiple methods.

where λ, β, and γ are hyperparameters controlling the strength of each regularization constraint.

A.5 More Cases for Case Study

Another case is shown in Figure 7.