CONFIDENCE FALLS SHORT: ASYMMETRIC CERTAINTY GAINS FROM OPTIMIZATION HINDER MULTIMODAL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning (MML) falls into the optimization dilemma due to the modality imbalance phenomenon, leading to suboptimal overall performance in practice. While many attempts primarily focus on balancing the optimization dynamics across modalities to address this issue, we identify a subtle yet critical flaw: optimization yields asymmetric gains in predictive certainty, with the strong modality more confident than the weak one, driving imbalanced modality contributions. In this paper, our analysis reveals that this flaw stems from unimodal characteristics rather than multimodal learning, and this confidence discrepancy can be corrected by positive cross-modal intervention. Based on this insight, we propose multimodal Max Confidence Regularization (MaxCR) to dynamically intervene in modality semantic confidence. Specifically, the semantic confidence of each modality is tracked using a nonlinear sparsity measure. We then design max suppression and max excitation based on this measure to regularize strong and weak modalities, respectively. They penalize and encourage the top-1 confidence, thereby constraining multimodal prediction. To this end, strong and weak modalities are expected to make calibrated confidence, thereby improving the overall performance. Empirical experiments on widely used datasets reveal the superiority of our method through comparison with various state-of-the-art (SOTA) multimodal learning baselines. The source code is available at `https://anonymous.4open.science/r/MaxCR-A07A`.

1 INTRODUCTION

Multimodal learning Baltrusaitis et al. (2019); Yang et al. (2022); Rakib & Bagavathi (2025); Lee et al. (2025); Tang et al. (2026); Liu et al. (2025) aims to effectively integrate heterogeneous information to improve the holistic understanding of tasks, which has received growing attention in various real applications Huang et al. (2024); Xiao et al. (2025); Hao et al. (2026). Despite expectations of achieving better performance compared with unimodal approaches, multimodal learning has been surprisingly shown to underperform compared to unimodal ones in certain scenarios Peng et al. (2022). The root of this problem lies in the existence of the modality imbalance phenomenon Wang et al. (2020), where different modalities converge at different rates.

Fortunately, many studies Wang et al. (2020); Peng et al. (2022); Zhang et al. (2024); Yang et al. (2024); Zong et al. (2024); Wei & Hu (2024); Wei et al. (2025) have explored this modality imbalance issue from various perspectives. As early pioneering studies, G-Blend Wang et al. (2020) and its successor OGM Peng et al. (2022), AGM Li et al. (2023), and PMR Fan et al. (2023), etc., focus on designing customized learning strategies for different modalities to intervene in their optimization processes to achieve rebalancing. Other attempts, including MLA Zhang et al. (2024), DI-MML Fan et al. (2024), and ReconBoost Hua et al. (2024), focus on decoupling multimodal learning and bridging the learning gap across modalities by injecting optimization information between them.

Despite significant progress, existing methods mainly adjust gradients to balance optimization dynamics, while overlooking unimodal intrinsic optimization characteristics. Compared with the strong modality, the weak modality often exhibits lower predictive confidence. We investigate this phenomenon on CREMA-D Cao et al. (2014). As shown in Figure 1 (left), we report accuracy, average top-1 confidence, and the average number of candidate classes with predicted probability

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

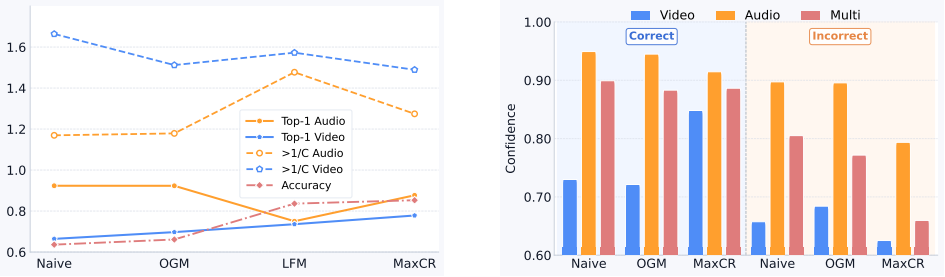


Figure 1: Comparison of the semantic confidence on CREMAD dataset. Left: We observe that the strong modality (Audio) exhibits higher Top-1 confidence and fewer candidate probabilities $> 1/C$, and the weak modality (Video) shows the opposite. Right: MaxCR mitigates underconfidence in correctly classified video samples and overconfidence in misclassified audio samples.

$> 1/C$, where C is the number of classes. Higher confidence and fewer candidates indicate more certain predictions. Under Naive MML and OGM Peng et al. (2022), substantial gaps exist between audio and video in both metrics. LFM Yang et al. (2024) introduces a contrastive loss Radford et al. (2021) to intervene in label fitting, potentially driving modalities toward flatter suboptimal basins while substantially improving overall performance, demonstrating the effectiveness of cross-modal supervision. As shown in Figure 1 (right), the confidence gap between strong and weak modalities appears on both correctly classified and misclassified samples, suggesting overconfidence in the strong modality and underconfidence in the weak one. In this paper, we analyze this problem through theory and experiments. We find that this discrepancy is not merely induced by multimodal learning, but fundamentally stems from unimodal characteristics. Specifically, for strong modality, the task gradient is positively correlated with the entropy reduction direction, causing predictions to become overly confident. In contrast, the weak modality is approximately orthogonal, resulting in smaller gains in predictive certainty. This asymmetric certainty gain already arises in unimodal learning and persists in multimodal learning. Ideally, each modality should produce confident but not overconfident predictions Zhou et al. (2025); Wei et al. (2022), enabling effective multimodal fusion.

Based on the above insight, we propose a novel multimodal Max Confidence Regularization (MaxCR) approach that leverages cross-modal information to calibrate the unimodal confidence, thereby improving the performance of decision-level multimodal fusion. Concretely, we first adopt a nonlinear sparsity measure to dynamically monitor the semantic confidence of each modality during training. Then, we design max suppression and max excitation based on this monitoring metric to regularize strong and weak modalities, respectively. The former penalizes the top-1 confidence in strong modality, while the latter encourages it in weak ones, thereby balancing multimodal fusion. To this end, MaxCR can calibrate and balance semantic confidence between strong and weak modalities, thereby improving multimodal performance. Extensive experiments on widely used datasets reveal that our proposed method MaxCR achieves competitive performance compared with SOTA baselines.

2 Related Work

2.1 Multimodal Learning

Multimodal learning Ngiam et al. (2011); Yang et al. (2019); Baltrusaitis et al. (2019) seeks to leverage these multimodal data from diverse sensors to improve task performance. Generally, existing multimodal learning methods can be categorized into early fusion Yang et al. (2022); Perez et al. (2018); Nie et al. (2021), late fusion Yang et al. (2019); Yao & Mihalcea (2022); Zhang et al. (2024), and hybrid fusion Zheng et al. (2023) based on their fusion strategies. In detail, early fusion methods typically perform joint modeling at the feature level, either by concatenating features to form a joint representation or by applying affine transformations. Representative early fusion methods include G-Blend Wang et al. (2020), DOMFN Yang et al. (2022), and so on. In contrast, late fusion methods tend to model different modalities separately, independently extracting predictions from modality-specific models and integrating them at the decision stage, such as MSLR Yao & Mihalcea (2022). Additionally, hybrid fusion methods aim to combine the advantages of early and late fusion,

typically introducing interaction mechanisms at different network layers to enable multi-level and multi-granularity cross-modal modeling Zheng et al. (2023). Despite numerous achievements, these methods may not fully exploit all modalities and produce under-optimized unimodal representations.

2.2 Imbalanced Multimodal Learning

Compared to unimodal approaches, multimodal learning is expected to achieve better performance Lv et al. (2021); Simonyan & Zisserman (2014); Gao et al. (2020). However, contrary to expectations, multimodal learning often encounters modality imbalance problems Wang et al. (2020); Peng et al. (2022) in practice, leading to performance degeneration and even underperformance compared to unimodal ones in certain scenarios. Imbalanced multimodal learning aims to rebalance the learning of different modalities, ensuring that each modality is sufficiently exploited. Based on this idea, early pioneering works Wang et al. (2020); Peng et al. (2022); Fan et al. (2023); Li et al. (2023) focus more on adaptively adjusting the optimization of different modalities. Representative methods, such as OGM Peng et al. (2022) and AGM Li et al. (2023), etc., adopt different learning strategies to modulate gradients or learning rates, thereby rebalancing the learning of weak and strong modalities. Other approaches, including MLA Zhang et al. (2024), ReconBoost Hua et al. (2024), and IGM Jiang et al. (2025a), take a different direction by focusing on decoupling multimodal learning and bridging the learning gap through injecting optimization information between modalities. In addition, several methods attempt to address the modality imbalance problem from different perspectives. Typically, LFM Yang et al. (2024) considers modality imbalance from a label-fitting perspective. AUG Jiang et al. (2025b) enhances the classification capability of weak modality to address modality imbalance.

3 Problem Analysis

In this section, we analyze the reason for the semantic confidence disparity between strong and weak modalities, and attribute it to asymmetric certainty gains from optimization. To bridge analysis and practice, we further provide empirical validation, confirming the existence of the optimization-certainty decoupling, which underlies the asymmetry.

3.1 Theoretical Analysis

For weak modality r, we assume that its Bayes posterior is $\boldsymbol{q}^r(\boldsymbol{x}^r) = \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X}^r = \boldsymbol{x}^r)$, where $\boldsymbol{x}^r$ is a realization of the random variable $\boldsymbol{X}^r$. Let $\boldsymbol{g}_{\mathrm{ce}} = \nabla_\theta \mathcal{L}_{\mathrm{CE}}$ denotes the population cross-entropy gradient, $\boldsymbol{g}_e = \nabla_\theta H$ denotes the entropy gradient, and $\hat{\boldsymbol{g}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{J}_i^\top (\boldsymbol{p}_i - \boldsymbol{y}_i)$ denote its mini-batch stochastic gradient, where N is the mini-batch size, $\boldsymbol{p}_i$ is the predicted class-probability vector for the i-th sample, $\boldsymbol{y}_i$ is its one-hot label vector, and $\boldsymbol{J}_i$ is the Jacobian of the model logits with respect to parameters θ. We further define the excess risk as $\mathcal{E}(\theta) = \mathbb{E}_{\boldsymbol{x}}\left[\mathrm{KL}(\boldsymbol{q}^r \| \boldsymbol{p})\right]$, and introduce the certainty-transfer ratio $r_N(\theta) = \frac{|\langle \boldsymbol{g}_e, \boldsymbol{g}_{\mathrm{ce}} \rangle|}{\|\boldsymbol{g}_e\| \sqrt{\mathbb{E}\|\hat{\boldsymbol{g}}\|^2}}$, which measures the proportion of predictable population cross-entropy gradient signal along the certainty-related direction in the stochastic update. $\mathrm{KL}(\cdot\|\cdot)$ is KL divergence. Our analysis builds on a common view: weak modality is typically difficult to model, prone to ambiguity, and even contaminated by noise Mai et al. (2024); Ding et al. (2025); Hwang et al. (2025); Gong et al. (2025). More details and proofs can be found in the appendix.

Theorem 1 (Optimization-Certainty Decoupling, Informal) *Under the assumptions that the weak modality is ambiguous, i.e., $\mathbb{E}_x[H(\boldsymbol{q}^r)] \geq H_w > 0$, where $H(\boldsymbol{q}^r)$ is the entropy and H_w is a constant representing the lower bound, and the model satisfies the regularity conditions. As $\mathcal{E}(\theta) \to 0$, $\|\boldsymbol{g}_{\mathrm{ce}}\| \leq B\sqrt{2\mathcal{E}(\theta)} \to 0$, $\mathbb{E}\|\hat{\boldsymbol{g}}\|^2 \geq \frac{\mu \phi_C(H_w)}{N} > 0$. Consequently, $r_N(\theta) \leq \frac{B\sqrt{2\mathcal{E}(\theta)}}{\sqrt{2B^2\mathcal{E}(\theta) + \mu\phi_C(H_w)/N}} \to 0$, $\dot{\mathcal{C}}(\theta) \to 0$, where B, N and μ are positive constants, and ϕ_C is a monotone function determined by the number of classes C.*

Theorem 1 shows that, as a weak modality approaches its Bayes-optimal cross-entropy solution, the population optimization signal vanishes, while irreducible conditional label ambiguity can sustain non-negligible stochastic gradient fluctuations. Consequently, substantial mini-batch gradients do not necessarily correspond to effective learning, since their magnitude may increasingly reflect ambiguity-induced stochastic variation rather than coherent population-level optimization. Meanwhile, the

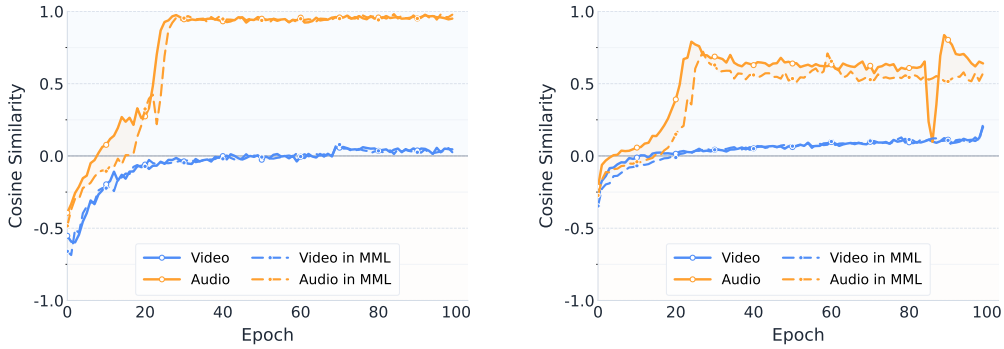


Figure 2: Cosine similarity analysis between the task gradient and the entropy-reduced gradient on CREMAD (left) and KSounds (right) datasets. In unimodal learning (Audio and Video), the task gradient and the entropy-reduced gradient exhibit high similarity for audio modality, i.e., strong modality, whereas they are orthogonal for video modality, i.e., weak modality. In multimodal learning (Audio in MML and Video in MML), this phenomenon persists, leading to disparity in confidence.

instantaneous certainty gain also approaches zero, indicating that continued parameter updates become progressively less effective at improving predictive certainty. This decoupling can further manifest as weak directional alignment between task optimization and certainty improvement, which we empirically examine in the following section.

3.2 Empirical Validation

Our theoretical analysis attributes the semantic confidence gap between strong and weak modalities to optimization-certainty decoupling. However, this framework relies on some simplifying assumptions to ensure analytical tractability. To bridge the gap between our idealized model and real-world behavior, we conduct empirical investigations. We investigate the directional alignment between task optimization and certainty improvement as an empirical indicator of inefficient certainty transfer.

Concretely, we update the model using the cross-entropy loss and evaluate the cosine similarity between the gradients induced by the unimodal entropy loss and the cross-entropy loss during training. We conduct experiments on CREMAD Cao et al. (2014) and KSounds Arandjelovic & Zisserman (2017) datasets. As shown in Figure 2, the results provide empirical evidence for a directional manifestation of optimization-certainty decoupling. We can observe that in unimodal learning, i.e., Audio and Video, the steepest descent direction of the audio modality is positively correlated with entropy reduction, whereas the cosine similarity of the video modality is close to zero, indicating that $r_N(\theta) \to 0$. In multimodal learning, i.e., Audio in MML and Video in MML, this phenomenon also persists. This phenomenon indicates that the strong modality naturally improves its prediction confidence during training, and the weak modality struggles to translate task optimization gains into enhanced prediction certainty. Experimental results on other datasets are provided in the appendix.

4 Methodology

4.1 Preliminaries

For simplicity, we use two modalities, i.e., audio and video, for illustration. It is worth mentioning that our method can be easily adapted to cases with more than two modalities. Notation definitions and the trimodal details are provided in the appendix.

Assume that we have M data points, each of which has audio and video modalities. Without loss of generality, we use $\mathcal{D} = \{\boldsymbol{x}_i^a, \boldsymbol{x}_i^v, \boldsymbol{y}_i\}_{i=1}^M$ to denote a training dataset, where $\boldsymbol{x}_i^a$ and $\boldsymbol{x}_i^v$ denote the i-th data point of audio and video description, respectively. In addition, we are also given a category label $\boldsymbol{y}_i \in \{0,1\}^C$ for each data point, where C denotes the number of category labels. With the training dataset $\mathcal{D}$, the objective of multimodal learning is to learn a model that fuses the multimodal information and predicts its category label as accurately as possible.

4.2 MULTIMODAL FRAMEWORK

For the sake of simplicity, we use superscript r to indicate the module corresponding to a specific modality, where $r \in \{a, v\}$. Following representative approaches Wang et al. (2020); Yang et al. (2024); Jiang et al. (2025b), we also utilize DNNs to construct our models. Specifically, we use $f^r(\cdot)$ to denote feature extractors, and the features can be calculated by $\boldsymbol{h}^r = f^r(\boldsymbol{x}^r; \theta^r)$, where θ^r denotes the extractor parameters. Then, the prediction of given data can be calculated by a classifier $\psi^r(\cdot)$: $\boldsymbol{p}^r = \psi^r(\boldsymbol{h}^r; \phi^r)$, where ϕ^r denotes the parameters of the classifier. Based on $\boldsymbol{p}^r$ and its ground-truth, the objective function can be written as:

$$\mathcal{L}_{CE}\left(\boldsymbol{x}^r, \boldsymbol{y}\right) = \frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{p}_i^r, \boldsymbol{y}_i) = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i^\top \log(\boldsymbol{p}_i^r), \tag{1}$$

where $\ell(\cdot)$ denotes the cross-entropy loss. Based on this, we decouple multimodal learning to address optimization conflicts Zhang et al. (2024); Hua et al. (2024). We then introduce cross-modal information to regularize modality-specific prediction confidence, thereby injecting optimization information across modalities and narrowing the confidence gap between them.

4.3 MAX CONFIDENCE REGULARIZATION

Based on the optimization-certainty decoupling phenomenon, we aim to address modality imbalance from the perspective of predictive confidence. Specifically, instead of directly reshaping optimization, we propose a novel approach, MaxCR, which dynamically identifies confidence discrepancy and regulates the maximum confidence of modality-specific predictions.

Confidence Monitoring. Since differences between modalities may evolve dynamically Peng et al. (2022), an adaptive score is required to estimate the semantic confidence of each modality. Therefore, we introduce a nonlinear sparsity measure to monitor them, which considers the semantic confidence and the candidate classes. Given a prediction $\boldsymbol{p}^r$, the sparsity metric can be formally defined as:

$$s_i^r = \frac{1}{L}\sum_{\boldsymbol{p}_{ij}^r \le \epsilon}^{L} exp(\frac{-\boldsymbol{p}_{ij}^r}{\tau}), \tag{2}$$

where $\tau \in (0, 1]$ scales the sparsity metric, and L denotes the number of entries smaller than ϵ. ϵ is the threshold for entries, with $\epsilon = \frac{1}{C}$. Based on this metric, we compute the sparsity score at t-th step on the fly and apply mild smoothing to mitigate early-stage fluctuations: $S_t^r = \frac{t-1}{t}\frac{1}{N}\sum_{i\in\boldsymbol{B}_t} s_i^r + \frac{1}{t}S_{t-1}^r$, where $\boldsymbol{B}_t$ is a mini-batch with size N and $S_0^r = 0$. A larger score indicates that the modality is more confident. Hence, if $S_t^a - S_t^v \ge \sigma$, we suppress the audio modality and excite the video modality, where σ is the dead zone for fault tolerance. Conversely, if $S_t^v - S_t^a \ge \sigma$, the video modality is suppressed and the audio modality is excited. We then design the discrepancy ratio λ:

$$\lambda = \left|S_t^a - S_t^v\right|. \tag{3}$$

This sparsity metric is more sensitive than entropy when predictions are uncertain and close to uniform, whereas entropy is more sensitive when the probability mass is concentrated on a few classes. Since label smoothing and entropy loss may over-amplify misclassified confidence Zhou et al. (2025); Zhang et al. (2025), we adopt a conservative peak regularization, ensuring the combined objective does not reverse the cross-entropy correction for an incorrect top-1 prediction. More details on the effectiveness and conservativeness are provided in the appendix.

Max Suppression. For over-confident modalities, we reduce excessive certainty. We formulate max suppression, which explicitly penalizes the top-1 confidence. Specifically, given a prediction $\boldsymbol{p}_i^r$, we compute the maximum probability and compare it with the mean. The loss is defined as follows:

$$\mathcal{L}_{sup}(\boldsymbol{x}^r, \boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} \lambda(\max_j \boldsymbol{p}_{ij}^r - \frac{1}{C}). \tag{4}$$

This loss reduces the top-1 confidence during optimization and encourages smooth predictions. In this way, max suppression mitigates the overconfidence of strong modality, thereby improving generalization. By combining (1) and (4), the strong modality objective can be defined as:

$$\mathcal{L}_s(\boldsymbol{x}^r, \boldsymbol{y}) = \mathcal{L}_{CE}(\boldsymbol{x}^r, \boldsymbol{y}) + \mathcal{L}_{sup}(\boldsymbol{x}^r, \boldsymbol{y}). \tag{5}$$

Algorithm 1 Learning algorithm of MaxCR.

Require: Training dataset $\mathcal{D} = \{\boldsymbol{x}_i^a, \boldsymbol{x}_i^v, \boldsymbol{y}_i\}_{i=1}^M$.
Ensure: The learned DNN models for all modalities.
INIT Initialize DNN parameters $\{\theta^r, \phi^r\}$,$r \in \{a, v\}$. Initialize iteration $t = 1$.
for $t = 1 \mapsto \#iterations$ **do**
  Sample a mini-batch $\boldsymbol{B}_t = \{\boldsymbol{x}_i^a, \boldsymbol{x}_i^v, \boldsymbol{y}_i\}_{i=1}^N$;
  $\forall \boldsymbol{x}_i^a, \boldsymbol{x}_i^v \in \boldsymbol{B}_t$, calculate predictions $\boldsymbol{p}_i^a$ and $\boldsymbol{p}_i^v$;
  Calculate sparsity score $\{S_t^a, S_t^v\}$ based on predictions; ▷ Max Confidence Regularization.
  Calculate discrepancy ratio λ in (3);
  if $S_t^a - S_t^v \geq \sigma$ **then** ▷ Max Suppression and Max Excitation.
    Calculate loss $\mathcal{L}_s(\boldsymbol{x}^a, \boldsymbol{y})$ and $\mathcal{L}_w(\boldsymbol{x}^v, \boldsymbol{y})$ in (5) and (7);
  else if $S_t^v - S_t^a > \sigma$ **then**
    Calculate loss $\mathcal{L}_s(\boldsymbol{x}^v, \boldsymbol{y})$ and $\mathcal{L}_w(\boldsymbol{x}^a, \boldsymbol{y})$ in (5) and (7);
  else
    Calculate loss in (1);
  end if
  Update DNN parameters $\{\theta^r, \phi^r\}$,$r \in \{a, v\}$;
end for

Max Excitation. For under-confident modalities, we encourage sharper predictions. We propose max excitation, which explicitly encourages the top-1 confidence. Similarly, given the prediction $\boldsymbol{p}_i^r$, we encourage maximum prediction to separate from the mean. The max excitation loss is defined as:

$$\mathcal{L}_{exc}(\boldsymbol{x}^r, \boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} \lambda(\frac{1}{C} - \max_j \boldsymbol{p}_{ij}^r). \tag{6}$$

This mechanism corrects predictions under high ambiguity, thereby alleviating underconfidence in weak modality. By combining (1) and (6), the weak modality objective can be defined as:

$$\mathcal{L}_w(\boldsymbol{x}^r, \boldsymbol{y}) = \mathcal{L}_{CE}(\boldsymbol{x}^r, \boldsymbol{y}) + \mathcal{L}_{exc}(\boldsymbol{x}^r, \boldsymbol{y}). \tag{7}$$

MaxCR is summarized in Algorithm 1. We obtain multimodal prediction by averaging the predictions across all modalities at inference. Combined with MaxCR, multimodal learning can be rebalanced.

5 Experiments

5.1 Experimental Setup

Dataset: CREMAD Cao et al. (2014) is an audio-visual dataset for emotion recognition, containing 7,442 video clips from 91 actors speaking several short words. **KSounds** Arandjelovic & Zisserman (2017) is designed for human action recognition, selected from Kinetics dataset, and contains 31 action classes. **VGGSound** Chen et al. (2020) is a large-scale audio-visual dataset in the wild with 309 classes and nearly 200k 10-second video clips. **Twitter** Yu & Jiang (2019) is an image-text dataset designed for emotion recognition, consisting of tweets collected from Twitter with three classes. **Sarcasm** Cai et al. (2019) dataset is designed for sarcasm detection, consisting of 24,635 image-text pairs. **NVGesture** Molchanov et al. (2016) is a trimodal dataset for gesture recognition, containing 1,532 dynamic gestures. More details are provided in the appendix.

Baselines: We select two categories of methods for comparison, i.e., traditional multimodal fusion methods and rebalanced multimodal learning methods. In detail, traditional multimodal fusion methods include concatenation fusion (Concat), affine transformation fusion (Affine) Perez et al. (2018), and multi-layer LSTM fusion (ML-LSTM) Nie et al. (2021). Rebalanced multimodal learning methods include G-Blend Wang et al. (2020), MSLR Yao & Mihalcea (2022), PMR Fan et al. (2023), AGM Li et al. (2023), MMPareto Wei & Hu (2024), SMV Wei et al. (2024), MLA Zhang et al. (2024), DI-MML Fan et al. (2024), ReconBoost Hua et al. (2024), LFM Yang et al. (2024), AMSS+ Yang et al. (2025), InfoReg Huang et al. (2025), AUG Jiang et al. (2025b).

Evaluation Protocols: We adopt accuracy, mean average precision (MAP), and MacroF1 as evaluation metrics. The accuracy measures the proportion of correct predictions of total predictions. MAP returns the average precision of all samples. MacroF1 calculates the average F1 across all categories.

Table 1: Comparison with SOTA multimodal learning methods. The best results are highlighted in **bold**. The underline denotes the second-best performance. The results with gray background denote the performance based on multimodal learning is inferior to that of the best unimodal approach.

Method	CREMAD		KSounds		VGGSound		Twitter		Sarcasm		NVGesture	
	Acc.	MAP	Acc.	MAP	Acc.	MAP	Acc.	F1	Acc.	F1	Acc.	F1
Unimodal-1	.6317	.5879	.5412	.5669	.4655	.4701	.5863	.4333	.7181	.7073	.7822	.7833
Unimodal-2	.4583	.6861	.5562	.5837	.3494	.3478	.7367	.6849	.8136	.8056	.7863	.7865
Unimodal-3	–	–	–	–	–	–	–	–	–	–	.8154	.8183
Concat	.6361	.6841	.6455	.7130	.5116	.5352	.7011	.6386	.8286	.8240	.8237	.8270
Affine	.6626	.7193	.6424	.6931	.5001	.5155	.7203	.5992	.8240	.8188	.8278	.8281
ML-LSTM	.6290	.6473	.6394	.6902	.4966	.5139	.7068	.6564	.8277	.8205	.8320	.8330
G-Blend	.6465	.7392	.6722	.7274	.5086	.5555	.7309	.6799	.8286	.8215	.8299	.8305
MSLR	.6868	.7412	.6756	.7282	.4987	.5415	.7232	.6382	.8439	.8378	.8237	.8284
OGM	.6612	.7372	.6582	.7159	.4829	.4978	.7058	.6435	.8360	.8293	–	–
PMR	.6659	.7058	.6675	.7274	.4647	.4866	.7357	.6636	.8310	.8256	–	–
AGM	.6733	.7807	.6791	.7388	.4711	.5198	.7261	.6502	.8306	.8293	.8279	.8284
MMPareto	.7487	.8535	.7000	.7850	.5125	.5473	.7358	.6729	.8348	.8284	.8382	.8424
SMV	.7872	.8417	.6900	.7426	.5031	.5362	.7428	.6817	.8418	.8368	.8352	.8341
MLA	.7943	.8572	.7004	.7945	.5165	.5473	.7352	.6713	.8426	.8348	.8340	.8372
DI-MML	.8158	.8592	.7203	.7426	.5173	.5479	.7248	.6686	.8411	.8315	–	–
ReconBoost	.7557	.8140	.6855	.7662	.5097	.5387	.7442	.6832	.8437	.8317	.8386	.8434
LFM	.8362	.9006	.7253	.7897	.5274	.5694	.7501	**.7057**	.8497	.8457	.8436	.8468
AMSS+	.7030	.7641	.7255	.7913	–	–	**.7558**	.6981	.8435	.8377	.8464	.8494
InfoReg	.7498	.8603	.7189	.7986	.5289	.5712	.7403	.6847	.8479	.8396	–	–
AUG	.8515	.9103	.7263	.7901	.5301	**.5826**	.7512	.6962	.8510	.8458	.8501	.8533
MaxCR	**.8531** ±0.0044	**.9216** ±0.0046	**.7522** ±0.0035	**.8102** ±0.0067	**.5416** ±0.0011	.5794 ±0.0032	.7517 ±0.0028	.6934 ±0.0043	**.8518** ±0.0021	**.8459** ±0.0016	**.8651** ±0.0062	**.8680** ±0.0043

Implementation Details: Following Peng et al. (2022), we employ ResNet18 He et al. (2016) as the backbone to encode audio and video for CREMAD, KSounds and VGGSound datasets. All the parameters of the backbone are randomly initialized. For NVGesture dataset, we employ the I3D Carreira & Zisserman (2017) as unimodal branch following the setting of Wu et al. (2022). We initialize the encoder with the pre-trained model trained on ImageNet. We use SGD with a momentum of 0.9 and a weight decay of 1e-4 as the optimizer. The learning rate is initialized at 1e-2 and decayed by a factor of 10 when the training loss reaches saturation for CREMAD, KSounds, VGGSound, and NVGesture datasets. For Twitter and Sarcasm datasets, following Yu & Jiang (2019); Cai et al. (2019), we adopt BERT Devlin et al. (2019) as the text encoder and ResNet50 He et al. (2016) as the image encoder. We use Adam Kingma & Ba (2015) as the optimizer, with an initial learning rate of 2e-5. For all datasets, τ is selected validation set from $\{\frac{1}{C}, \frac{1}{2C}, \frac{1}{3C}, \frac{1}{4C}, \frac{1}{5C}\}$ with the default set to $\frac{1}{C}$, while σ is set to be 0.1, where C is the number of categories. For comparison methods, the source codes of all baselines are kindly provided by their authors, and all baselines adopt the same backbone and initialization strategy. We conduct experiments with three random seeds. All experiments are conducted on NVIDIA GeForce RTX 4090, and all models are implemented with PyTorch.

5.2 Main Results

Comparison with Multimodal Baseline: The main results on all datasets are reported in Table 1. Unimodal-1/2 respectively denote the audio/video for CREMAD and KSounds, and image/text for Twitter and Sarcasm. Unimodal-1/2/3 denotes the RGB/OF/Depth modality for NVGesture dataset, respectively. Furthermore, the results with a gray background indicate that the performance based on multimodal learning is inferior to that of the best unimodal approach. First, the results show that our method consistently achieves competitive performance on all datasets. In contrast, compared with direct unimodal learning, some multimodal baseline methods fail to yield improvements and even produce worse results. Moreover, results on NVGesture dataset demonstrate that MaxCR effectively addresses challenges in scenarios involving more than two modalities and achieves best performance.

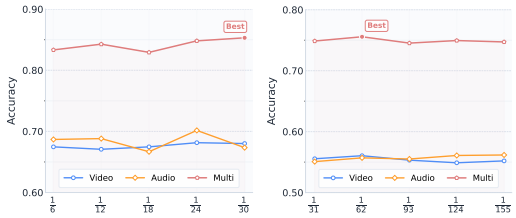


Figure 3: Sensitivity analysis about τ on CREMAD (left) and KSounds (right) datasets.

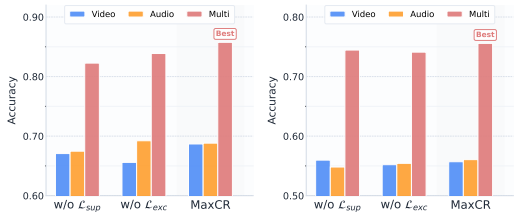


Figure 4: Ablation study of MaxCR on CREMAD (left) and KSounds (right) datasets.

Table 2: Performance with different calibration metrics on KSounds dataset.

Method	Acc	ECE	NLL	Brier
Naive	0.6455	0.1607	1.5206	0.5167
OGM	0.6582	0.1498	1.4561	0.5009
MaxCR	**0.7522**	**0.0384**	**0.9195**	**0.3466**

Table 3: Performance with different monitoring metrics on KSounds dataset.

Method	Audio	Video	Multimodal
Confidence	0.5322	0.5354	0.7317
Entropy	0.5385	0.5443	0.7352
Sparsity (Our)	**0.5570**	**0.5605**	**0.7522**

5.3 Sensitivity and Ablation Study

Sensitivity to the scaling factor τ: We explore the influence of scaling factor τ on CREMAD and KSounds datasets. We present the accuracy with different $\tau \in \left[\frac{1}{5C}, \frac{1}{C}\right]$ in Figure 3. Although the optimal τ varies across datasets, MaxCR remains insensitive to the scaling factor τ in a large range.

Ablation Study: We analyzed the impact of losses $\mathcal{L}_{sup}$ and $\mathcal{L}_{exc}$, i.e., Equation (4) and (6), on CREMAD and KSounds datasets to investigate the effectiveness of MaxCR. As shown in Figure 4, we can observe that when the max suppression loss is removed (w/o $\mathcal{L}_{sup}$), the accuracy of the strong modality, i.e., Audio, decreases. Conversely, when the max excitation loss is removed (w/o $\mathcal{L}_{exc}$), the accuracy of the weak modality, i.e., Video, declines. When both max suppression and max excitation are used together, the two modalities achieve balance, leading to the best multimodal performance.

5.4 Further Analysis

Calibration Metrics Evaluation: We incorporate the Expected Calibration Error (ECE), Negative Log-Likelihood (NLL), and Brier Score (Brier) metrics to evaluate confidence. Fundamentally, ECE quantifies the discrepancy between predictive confidence and empirical accuracy. NLL measures the probability assigned to the ground-truth label, penalizing confident mispredictions, and the Brier Score captures the squared deviation between predicted probabilities and the true label distribution. We conduct experiments on KSounds dataset. As reported in Table 2, the results compellingly demonstrate that, beyond performance gains, MaxCR yields more reliable semantic confidence.

Selection of Monitoring Metric: We introduce a sparsity metric to quantify the confidence discrepancy across modalities. To some extent, entropy and label confidence can also serve as monitoring metrics. To evaluate the effectiveness of entropy and confidence, we conduct experiments on KSounds dataset. Specifically, we use normalized entropy and the average label confidence to ensure the same numerical scale. As shown in Table 3, the entropy achieves better performance than confidence, while the sparsity metric achieves the best accuracy. More results are provided in the appendix.

Max Suppression vs. Label Smoothing: Since max suppression loss is designed to mitigate overconfidence, it naturally relates to label smoothing, which combines a uniform distribution with hard ground-truth labels to alleviate overconfidence. We experiment on KSounds dataset to examine whether label smoothing yields gains. Specifically, label smoothing is dynamically applied to the strong modality during training. The results in Table 4 show that the improvement brought by label smoothing (w/ LS) is weaker than that of the proposed max suppression (w/ Sup) strategy. Max suppression improves the performance of the audio modality and enhances its collaboration with the video modality. This may be because max suppression regularizes the peaks of model predictions, rather than uniformly suppressing the ground-truth class.

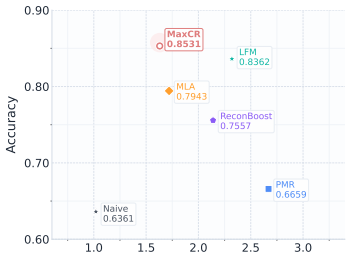


Figure 5: Training time (hrs).

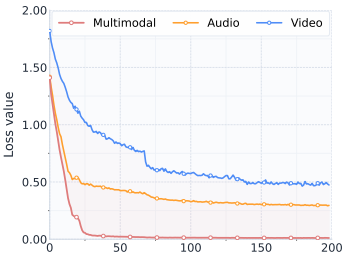


Figure 6: Loss visualization.

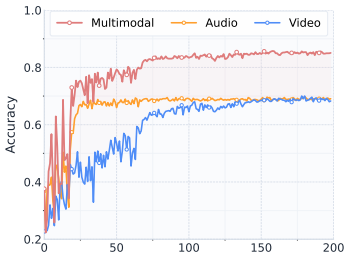


Figure 7: Accuracy visualization.

Table 4: Comparison of label smoothing and max suppression on KSounds dataset.

Method	Audio	Video	Multimodal
w/ LS	0.5466	0.5226	0.7286
w/ Sup	0.5543	0.5520	0.7410
MaxCR	**0.5570**	**0.5605**	**0.7522**

Table 5: Comparison of entropy minimization and max excitation on KSounds dataset.

Method	Audio	Video	Multimodal
w/ EM	0.5474	0.5280	0.7248
w/ Exc	0.5481	0.5597	0.7445
MaxCR	**0.5570**	**0.5605**	**0.7522**

Max Excitation vs. Entropy Minimization: Since max excitation loss encourages more confident predictions, its objective is closely related to that of entropy minimization. Therefore, we further compare the max excitation loss with the entropy minimization loss on KSounds dataset. Specifically, we apply entropy minimization loss to the weak modality. As shown in Table 5, the gain from entropy minimization (w/ EM) is smaller than that from max excitation (w/ Exc).

Training Overhead: We empirically compare the training overhead of MaxCR with that of competing multimodal learning baselines under the same experimental setting, including Naive, PMR, MLA, ReconBoost, and LFM. The results are shown in Figure 5, where training times are reported in hours. It can be observed that MaxCR achieves the best accuracy with limited training time.

Convergence: In Figure 6, we present the convergence behavior of MaxCR on CREMAD dataset during training. It can be observed that the loss gradually converges as training proceeds without getting stuck. In addition, Figure 7 reports the accuracy throughout training. We can observe that the accuracy exhibits a similar convergence trend.

6 Conclusion

In this paper, we revisit modality imbalance in multimodal classification from the perspective of predictive certainty. We show that optimization yields asymmetric certainty gains: strong modalities become increasingly confident, whereas weak modalities struggle to translate optimization progress into predictive certainty. We characterize this behavior as optimization-certainty decoupling, which emerges in unimodal learning and persists in multimodal learning. Motivated by this insight, we propose multimodal Max Confidence Regularization (MaxCR), which dynamically regulates modality-specific predictions. Specifically, MaxCR uses a nonlinear sparsity measure to monitor semantic confidence and applies max suppression and max excitation to regularize the top-1 confidence of strong and weak modalities, respectively. To this end, MaxCR encourages confident yet not overconfident predictions, improving multimodal fusion. Experiments on widely used datasets demonstrate the effectiveness of MaxCR. Overall, our results highlight predictive certainty as an important yet underexplored dimension of modality imbalance and establish confidence-aware intervention as a simple yet effective complement to balance optimization strategies.

Limitations: Our proposed method MaxCR focuses on multimodal classification. Extending it to other multimodal tasks may require task-specific confidence measures and regularization objectives. In addition, our theoretical analysis provides a tractable characterization of optimization-certainty decoupling, while a more general treatment of certainty evolution throughout optimization warrants further investigation. We leave a more comprehensive investigation as future work.

AI USE STATEMENT

We used generative AI tools solely for language proofreading to improve the clarity and readability of the manuscript and did not use them for other parts. All scientific content, including ideas, methods, and experiments, etc., was generated and verified by the authors.

REFERENCES

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pp. 609–617. IEEE, 2017.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 41(2):423–443, 2019.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, pp. 2506–2515, 2019.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: crowd-sourced emotional multimodal actors dataset. *TAC*, 5(4):377–390, 2014.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pp. 4724–4733. IEEE, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pp. 721–725. IEEE, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Xiaorui Ding, Huan Ma, and Changqing Zhang. A theoretical proof of dynamic multimodal fusion exacerbates modality greedy. In *ACM MM*, pp. 2429–2436. ACM, 2025.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: prototypical modal rebalance for multimodal learning. In *CVPR*, pp. 20029–20038. IEEE, 2023.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. Detached and interactive multimodal learning. In *ACM MM*. ACM, 2024.

Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, pp. 10454–10464. IEEE, 2020.

Baoquan Gong, Xiyuan Gao, Pengfei Zhu, Qinghua Hu, and Bing Cao. Multimodal negative learning. In *NeurIPS*, 2025.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. Uni-x: Mitigating modality conflict with a two-end-separated architecture for unified multimodal models. In *ICLR*. OpenReview.net, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778. IEEE, 2016.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. In *ICML*. PMLR, 2024.

Chengxiang Huang, Yake Wei, Zequn Yang, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition. In *CVPR*, pp. 25854–25863. IEEE, 2025.

Longfei Huang, Xiangyu Wu, Jingyuan Wang, Weili Guo, and Yang Yang. Refining visual perception for decoration display: A self-enhanced deep captioning model. In *ACML*, 2024.

Seonghyeon Hwang, Soyoung Choi, and Steven Euijong Whang. MIDAS: misalignment-based data augmentation strategy for imbalanced multimodal learning. In *NeurIPS*, 2025.

Qing-Yuan Jiang, Zhouyang Chi, and Yang Yang. Interactive multimodal learning via flat gradient modification. In *IJCAI*, pp. 5489–5497. ijcai.org, 2025a.

Qing-Yuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. In *NeurIPS*, 2025b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*. OpenReview.net, 2015.

Jungsoo Lee, Janghoon Cho, Hyojin Park, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. Generalized contrastive learning for universal multimodal retrieval. In *NeurIPS*, 2025.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pp. 22157–22167. IEEE, 2023.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. In *NeurIPS*, 2025.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *CVPR*, pp. 2554–2562. IEEE, 2021.

Sijie Mai, Ya Sun, Aolin Xiong, Ying Zeng, and Haifeng Hu. Multimodal boosting: Addressing noisy modalities and identifying modality contribution. *TMM*, 26:3018–3033, 2024.

Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *CVPR*, pp. 4207–4215. IEEE, 2016.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pp. 689–696. PMLR, 2011.

Weizhi Nie, Yan Yan, Dan Song, and Kun Wang. Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition. *MTA*, 80(11):16205–16214, 2021.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pp. 8228–8237. IEEE, 2022.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, pp. 3942–3951. AAAI Press, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Mohammed Rakib and Arunkumar Bagavathi. G^2d: Boosting multimodal learning with gradient-guided distillation. In *ICCV*, pp. 4059–4068. IEEE, 2025.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

Zhijiang Tang, Linhua Wang, Jiaxin Qi, Weihao Jiang, Peng Hou, Anxiang Zeng, and Jianqiang Huang. Cccaption: Dual-reward reinforcement learning for complete and correct image captioning. In *CVPR*. IEEE, 2026.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pp. 12692–12702. IEEE, 2020.

Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu. To smooth or not? when label smoothing meets noisy labels. In *ICML*, pp. 23589–23614. PMLR, 2022.

Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, pp. 52559–52572. PMLR, 2024.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, pp. 27338–27347. IEEE, 2024.

Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *TPAMI*, 47(1):469–485, 2025.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, pp. 24043–24055. PMLR, 2022.

Jian Xiao, Zijie Song, Jialong Hu, Hao Cheng, Jia Li, Zhenzhen Hu, and Richang Hong. Rebalancing contrastive alignment with bottlenecked semantic increments in text-video retrieval. In *NeurIPS*, 2025.

Yang Yang, Ke-Tao Wang, De-Chuan Zhan, Hui Xiong, and Yuan Jiang. Comprehensive semi-supervised multi-modal learning. In *IJCAI*, pp. 4092–4098. ijcai, 2019.

Yang Yang, Jingshuai Zhang, Fan Gao, Xiaoru Gao, and Hengshu Zhu. DOMFN: A divergence-orientated multi-modal fusion network for resume assessment. In *ACM MM*, pp. 1612–1620. ACM, 2022.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *NeurIPS*, 2024.

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *TPAMI*, 47(6):4553–4566, 2025.

Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *ACL*, pp. 1824–1834, 2022.

Jianfei Yu and Jing Jiang. Adapting BERT for target-oriented multimodal sentiment classification. In *IJCAI*, pp. 5408–5414. ijcai.org, 2019.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: test-time adaption by conservatively minimizing entropy. In *ICLR*. OpenReview.net, 2025.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *CVPR*, pp. 27456–27466. IEEE, 2024.

Xiao Zheng, Chang Tang, Zhiguo Wan, Chengyu Hu, and Wei Zhang. Multi-level confidence learning for trustworthy multimodal classification. In *AAAI*, pp. 11381–11389. AAAI Press, 2023.

Yuxuan Zhou, Heng Li, Zhi-Qi Cheng, Xudong Yan, Yifei Dong, Mario Fritz, and Margret Keuper. Maxsup: Overcoming representation collapse in label smoothing. In *NeurIPS*, 2025.

Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, and Ken Zheng. Balancing multimodal learning via online logit modulation. In *IJCAI*, pp. 5753–5761. ijcai.org, 2024.

APPENDIX CONTENTS

A Theoretical Details and Proofs **14**
A.1 Preliminaries and Assumptions 14
A.2 Auxiliary Lemmas and Proofs 14
A.3 Theorems, Proofs and Corollaries 15

B Analysis of MaxCR **18**
B.1 Certainty Effect 18
B.2 Comparison with Entropy and Conservativeness 18

C Notation Definition **19**

D Experiment Details **20**
D.1 Dataset Details 20
D.2 More Implementation Details 20
D.3 Extension to More than Two Modalities 20

E Additional Experiments **20**
E.1 Further Validation Experiments 20
E.2 Selection of Monitoring Metric on CREMAD dataset 21
E.3 Incorporating Optimization Dynamics 21
E.4 Analysis on Semantic Confidence 22
E.5 Visualization Results 22

A Theoretical Details and Proofs

A.1 Preliminaries and Assumptions

For clarity, we consider two modalities. Let the parameters of the r-th modality be $\theta^r \in \mathbb{R}^d$. The network produces logits $\boldsymbol{z}(\boldsymbol{x}^r;\theta) \in \mathbb{R}^C$, with predictions $\boldsymbol{p} = \mathrm{Softmax}(\boldsymbol{z})$. The Jacobian is defined as $\boldsymbol{J}_\theta(x^r) = \partial \boldsymbol{z}/\partial\theta \in \mathbb{R}^{C\times d}$, and the sample-wise neural tangent kernel (NTK) is $K_\theta(\boldsymbol{x}^r, \boldsymbol{x}^{r\prime}) = \boldsymbol{J}_\theta(\boldsymbol{x}^r)J_\theta(\boldsymbol{x}^{r\prime})^\top$. Define the zero-sum subspace as $\mathcal{S}_0 = \{\boldsymbol{v} \in \mathbb{R}^C : \mathbf{1}^\top \boldsymbol{v} = 0\}$, $\Pi_0 = \boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$. We further define $\boldsymbol{g}_{ce} = \nabla_\theta \mathcal{L}_{CE}$, $\boldsymbol{g}_e = \nabla_\theta H$ and certainty $\mathcal{C}(\theta) \triangleq -\mathbb{E}_x[H(\boldsymbol{p})]$.

Assumption 1 (Modality Ambiguity) *Let $\boldsymbol{q}^r(\boldsymbol{x}^r) = \mathbb{P}(\boldsymbol{y} \mid \boldsymbol{x}^r) \in \Delta^{C-1}$ denote the Bayes posterior of r-modality. There exist constants $H_s < H_w$ such that $\mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q}^a)] \le H_s$, $\mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q}^v)] \ge H_w$. Equivalently, $H(\boldsymbol{Y} \mid \boldsymbol{X}^a) \ll H(\boldsymbol{Y} \mid \boldsymbol{X}^v)$, i.e., the strong modality induces nearly deterministic labels, whereas the labels conditioned on the weak modality are intrinsically ambiguous.*

Assumption 2 (Regularity) *There exist constants B, μ such that, for all x, $|\boldsymbol{J}_\theta(\boldsymbol{x})|2 \le B$, and the kernel is well-conditioned on $\mathcal{S}_0$: $\mu\Pi_0 \preceq \Pi_0 K\theta(x,x)\Pi_0 \preceq L\Pi_0$, $\mu > 0$.*

The population cross-entropy loss admits the decomposition $\mathcal{L}_{CE}(\theta) = \mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q})] + \mathbb{E}_{\boldsymbol{x}}[\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p})]$. We define the excess risk as $\mathcal{E}(\theta) = \mathbb{E}_x[\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p})]$.

Under gradient flow $\dot\theta = -\eta \boldsymbol{g}_{ce}$, we have the identity

$$\dot{\mathcal{C}}(\theta) = \eta\langle \boldsymbol{g}_e, \boldsymbol{g}_{ce}\rangle. \tag{8}$$

Let $\boldsymbol{a}(\boldsymbol{p}) = \log \boldsymbol{p} + H(\boldsymbol{p})\mathbf{1}$, which satisfies $\langle \boldsymbol{p}, \boldsymbol{a}\rangle = 0$. Since $\nabla_{\boldsymbol{z}} H = -\boldsymbol{p} \odot \boldsymbol{a}$ and $\nabla_{\boldsymbol{z}} \ell = \boldsymbol{p} - \boldsymbol{y}$, define

$$\Psi(\boldsymbol{p}, \boldsymbol{y}) \triangleq \langle \nabla_{\boldsymbol{z}} H, \nabla_{\boldsymbol{z}} \ell\rangle = \boldsymbol{p}_y \boldsymbol{a}_y - \Gamma(\boldsymbol{p}), \tag{9}$$

where $\Gamma(\boldsymbol{p}) = \sum_j \boldsymbol{p}_j^2 \boldsymbol{a}_j = \mathrm{Cov}_{j\sim\boldsymbol{p}}(\boldsymbol{p}_j, \log \boldsymbol{p}_j) \ge 0$. Ψ is the sample-wise certainty transfer coefficient. Under an isotropic kernel, $\dot{\mathcal{C}} = \eta\mathbb{E}[\Psi]$. Its sign is controlled by a candidate-set criterion determined solely by the predictive distribution.

Definition 1 (Candidate Class Set) $\mathbb{C}(\boldsymbol{p}) = \{j : \boldsymbol{p}_j > e^{-H(\boldsymbol{p})}\}$. *We always have* $\arg\max_j \boldsymbol{p}_j \in \mathbb{C}(\boldsymbol{p})$.

The mini-batch gradient $\hat{\boldsymbol{g}} = \frac{1}{N}\sum_i \boldsymbol{J}_i^\top(\boldsymbol{p}_i - \boldsymbol{y}_i)$ satisfies $\mathbb{E}\hat{\boldsymbol{g}} = \boldsymbol{g}_{ce}$ and $\mathbb{E}\|\hat{\boldsymbol{g}}\|^2 = \|\boldsymbol{g}_{ce}\|^2 + \frac{1}{N}\mathrm{tr}\Sigma$. To handle arbitrary $C \ge 2$ in a unified manner, define

$$F_C(e) \triangleq H_b(e) + e\log(C-1), e \in [0, 1-1/C], \tag{10}$$

and its inverse

$$\phi_C \triangleq F_C^{-1} : [0, \log C] \to [0, 1-1/C]. \tag{11}$$

Over this interval, F_C is monotonically increasing and concave, and hence ϕ_C is monotonically increasing and convex.

We further define the effective certainty transfer rate:

$$r_N(\theta) \triangleq \frac{|\langle \boldsymbol{g}_e, \boldsymbol{g}_{ce}\rangle|}{\|\boldsymbol{g}_e\|\sqrt{\mathbb{E}\|\hat{\boldsymbol{g}}\|^2}}. \tag{12}$$

A.2 Auxiliary Lemmas and Proofs

We have the following basic identity:

Lemma 1 *We have $\nabla_{\boldsymbol{z}} H(\boldsymbol{p}) = -\boldsymbol{p} \odot \boldsymbol{a}$, where $\boldsymbol{a} = \log \boldsymbol{p} + H(\boldsymbol{p})\mathbf{1}$. Moreover, $\langle \boldsymbol{p}, \boldsymbol{a}\rangle = 0$, $\mathbf{1}^\top \nabla_{\boldsymbol{z}} H = 0$.*

Proof 1 *Using $\partial \boldsymbol{p}_j / \partial \boldsymbol{z}_c = \boldsymbol{p}_j(\delta_{jc} - \boldsymbol{p}_c)$, we obtain*

$$\begin{aligned} \frac{\partial H}{\partial \boldsymbol{z}_c} &= -\sum_j \frac{\partial \boldsymbol{p}_j}{\partial \boldsymbol{z}_c}(\log \boldsymbol{p}_j + 1) \\ &= -\boldsymbol{p}_c(\log \boldsymbol{p}_c + 1) + \boldsymbol{p}_c \sum_j \boldsymbol{p}_j(\log \boldsymbol{p}_j + 1) \\ &= -\boldsymbol{p}_c(\log \boldsymbol{p}_c + H). \end{aligned} \tag{13}$$

Hence, $\nabla_{\boldsymbol{z}} H(\boldsymbol{p}) = -\boldsymbol{p} \odot \boldsymbol{a}$. Furthermore, $\langle \boldsymbol{p}, \boldsymbol{a} \rangle = \sum_j \boldsymbol{p}_j \log \boldsymbol{p}_j + H = 0$, which implies $\mathbf{1}^\top \nabla_{\boldsymbol{z}} H = -\langle \boldsymbol{p}, \boldsymbol{a} \rangle = 0$.

This completes the proof.

Lemma 2 *$\Gamma(\boldsymbol{p}) = \sum_j \boldsymbol{p}_j^2 \boldsymbol{a}_j = \mathrm{Cov}_{j \sim \boldsymbol{p}}(\boldsymbol{p}_j, \log \boldsymbol{p}_j) \geq 0$, with equality if and only if $\boldsymbol{p}$ is uniform over its support.*

Proof 2 *Let $j \sim \boldsymbol{p}$. Since $\mathbb{E}_{\boldsymbol{p}}[\log \boldsymbol{p}_j] = -H$ and $\mathbb{E}_{\boldsymbol{p}}[\boldsymbol{p}_j] = \|\boldsymbol{p}\|_2^2$, we have $\mathrm{Cov}(\boldsymbol{p}_j, \log \boldsymbol{p}_j) = \sum_j \boldsymbol{p}_j^2 \log \boldsymbol{p}_j + H\|\boldsymbol{p}\|_2^2 = \sum_j \boldsymbol{p}_j^2 \boldsymbol{a}_j = \Gamma$.*

Since larger $\boldsymbol{p}_j$ always corresponds to larger $\log \boldsymbol{p}_j$, the Chebyshev association inequality (FKG) gives a nonnegative covariance. Equality requires $\boldsymbol{p}_j$ to be almost surely constant.

This completes the proof.

Lemma 3 *$e^{-H(\boldsymbol{p})} \leq \max_j \boldsymbol{p}_j$. Hence, $\mathbb{C}(\boldsymbol{p}) \neq \varnothing$ and contains the top-1 class.*

Proof 3 *Since $H = \sum_j \boldsymbol{p}_j \log(1/\boldsymbol{p}_j) \geq \log(1/\boldsymbol{p}_{\max})$, the result follows.*

This completes the proof.

A.3 Theorems, Proofs and Corollaries

Theorem 1 (Certainty upper bound) *Under Assumption 1, for any θ, let $\bar{\varepsilon} = \sqrt{\mathcal{E}(\theta)/2} \leq 1 - 1/C$. Then $\mathcal{C}(\theta) \leq -\mathbb{E}_x[H(\boldsymbol{q})] + \bar{\varepsilon} \log(C-1) + H_b(\bar{\varepsilon})$ where H_b denotes the binary entropy function. In particular, any population minimizer of the cross-entropy loss satisfies $\mathcal{C}(\theta^\star) = -\mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q})]$.*

Proof 4 *Pointwise,*

$$\mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{q}}[-\log \boldsymbol{p}_y] = H(\boldsymbol{q}) + \mathrm{KL}(\boldsymbol{q} \| \boldsymbol{p}). \tag{14}$$

Taking the expectation over $\boldsymbol{x}$ gives $\mathcal{L}_{CE} = \mathbb{E}[H(\boldsymbol{q})] + \mathcal{E}(\theta)$. Let $T(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{p}(\boldsymbol{x}) - \boldsymbol{q}(\boldsymbol{x})\|_1$. By Pinsker's inequality,

$$T \leq \sqrt{\mathrm{KL}(\boldsymbol{q} \| \boldsymbol{p})/2}. \tag{15}$$

By the Fannes–Audenaert continuity bound for entropy, when $T \leq 1 - \frac{1}{C}$, we have

$$|H(\boldsymbol{p}) - H(\boldsymbol{q})| \leq T \log(C-1) + H_b(T). \tag{16}$$

Taking the expectation over x, and noting that $t \mapsto t \log(C-1) + H_b(t)$ is concave on $[0, 1 - 1/C]$, Jensen's inequality together with

$$\mathbb{E}[T] \leq \sqrt{\mathbb{E}[\mathrm{KL}]/2} = \bar{\varepsilon}, \tag{17}$$

(Cauchy–Schwarz + Jensen) yields

$$\mathbb{E}[H(\boldsymbol{q})] - \mathbb{E}[H(\boldsymbol{p})] \leq \bar{\varepsilon} \log(C-1) + H_b(\bar{\varepsilon}). \tag{18}$$

Rearranging the terms gives the result. If θ is a population minimizer and the model is realizable, then $\boldsymbol{p} = \boldsymbol{q}$ almost everywhere, and therefore

$$\mathcal{C}(\theta^\star) = -\mathbb{E}[H(\boldsymbol{q})]. \tag{19}$$

This completes the proof.

Corollary 1 (Asymmetric Gain) *Under Assumption 1, the certainty attainable under cross-entropy training satisfies $\mathcal{C}^{\boldsymbol{a}} \geq -H_s - o(1)$ and $\mathcal{C}^{v} \leq -H_w + o(1)$. Hence, the confidence gap is at least $H_w - H_s$, and this gap is independent of the optimization algorithm or gradient reweighting scheme.*

Theorem 2 (Certainty Gain) *If $\mathbb{C}(\boldsymbol{p}) = \{\boldsymbol{y}\}$, i.e., the prediction is correct and the candidate class is unique, then $\Psi(\boldsymbol{p}, \boldsymbol{y}) \geq \boldsymbol{p}_y(1 - \boldsymbol{p}_y)\boldsymbol{a}_y > 0$. If $\boldsymbol{y} \notin \mathbb{C}(\boldsymbol{p})$, then $\Psi(\boldsymbol{p}, \boldsymbol{y}) \leq -\Gamma(\boldsymbol{p}) \leq 0$. If the label follows an ambiguous posterior $y \sim \boldsymbol{q}$, then $\mathbb{E}_{y\sim\boldsymbol{q}}\Psi = \langle \boldsymbol{q} - \boldsymbol{p}, \boldsymbol{p} \odot \boldsymbol{a}\rangle$.In particular, when $\boldsymbol{q}$ is uniform, $\mathbb{E}_{y\sim\boldsymbol{q}}\Psi = -\Gamma(\boldsymbol{p}) \leq 0$.*

Proof 5 *By Lemma 1,*

$$\Psi = \langle -\boldsymbol{p} \odot \boldsymbol{a}, \boldsymbol{p} - \boldsymbol{y}\rangle = \boldsymbol{p}_y\boldsymbol{a}_y - \Gamma(\boldsymbol{p}). \tag{20}$$

Suppose $\mathbb{C}(\boldsymbol{p}) = \{\boldsymbol{y}\}$. Then $\boldsymbol{a}_y = \log \boldsymbol{p}_y + H \geq 0$. For any $j \neq y$,

$$\boldsymbol{p}_j \leq e^{-H} \Rightarrow \boldsymbol{a}_j \leq 0 \Rightarrow \boldsymbol{p}_j^2\boldsymbol{a}_j \leq 0. \tag{21}$$

Therefore,

$$\Gamma = \boldsymbol{p}_y^2\boldsymbol{a}_y + \sum_{j\neq y} \boldsymbol{p}_j^2\boldsymbol{a}_j \leq \boldsymbol{p}_y^2\boldsymbol{a}_y, \tag{22}$$

and hence

$$\Psi \geq \boldsymbol{p}_y\boldsymbol{a}_y - \boldsymbol{p}_y^2\boldsymbol{a}_y = \boldsymbol{p}_y(1 - \boldsymbol{p}_y)\boldsymbol{a}_y. \tag{23}$$

By Lemma 3, $\boldsymbol{p}_y = \boldsymbol{p}_{\max}$. If $\boldsymbol{p}$ is non-uniform, then $\boldsymbol{a}_y > 0$, and the inequality is strict.

If $\boldsymbol{y} \notin \mathbb{C}(\boldsymbol{p})$, then $\boldsymbol{a}_y \leq 0$, so $\boldsymbol{p}_y\boldsymbol{a}_y \leq 0$. By Lemma 2, $\Gamma \geq 0$, and therefore

$$\Psi \leq -\Gamma \leq 0. \tag{24}$$

We have

$$\mathbb{E}_{y\sim\boldsymbol{q}}\Psi = \sum_j \boldsymbol{q}_j\boldsymbol{p}_j\boldsymbol{a}_j - \Gamma = \langle \boldsymbol{q}, \boldsymbol{p} \odot \boldsymbol{a}\rangle - \langle \boldsymbol{p}, \boldsymbol{p} \odot \boldsymbol{a}\rangle = \langle \boldsymbol{q} - \boldsymbol{p},\ \boldsymbol{p} \odot \boldsymbol{a}\rangle. \tag{25}$$

For $\boldsymbol{q} = \boldsymbol{u} = \mathbf{1}/C$,

$$\langle \boldsymbol{u}, \boldsymbol{p} \odot \boldsymbol{a}\rangle = \frac{1}{C}\langle \mathbf{1}, \boldsymbol{p} \odot \boldsymbol{a}\rangle = 0, \tag{26}$$

so

$$\mathbb{E}\Psi = -\Gamma \leq 0. \tag{27}$$

This completes the proof.

Corollary 2 (Stability) *For a general posterior $\boldsymbol{q}$, $\mathbb{E}_{y\sim\boldsymbol{q}}\Psi \leq -\Gamma(\boldsymbol{p}) + \omega(\boldsymbol{p})\|\boldsymbol{q} - \boldsymbol{u}\|_1$, where $\omega(\boldsymbol{p}) = \max_j \boldsymbol{p}_j|\boldsymbol{a}_j| \leq e^{-1} + \log C$. Thus, as the posterior approaches the uniform distribution, the certainty gain is increasingly driven toward non-positive values.*

Theorem 3 (Optimization-Certainty Decoupling) *Under Assumptions 1 and 2, we have $\mathrm{tr}\Sigma \geq \mu\mathbb{E}_{\boldsymbol{x}}\left[1 - \|\boldsymbol{q}\|_2^2\right] \geq \mu\phi_C\left(\mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q})]\right)$ and $\|\boldsymbol{g}_{ce}\| \leq B\sqrt{2\mathcal{E}(\theta)}$. Therefore, for the weak modality, if $\mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q}^v)] \geq H_w > 0$, then, as $\mathcal{E}(\theta) \to 0$, $\dot{\mathcal{C}} \to 0, \mathbb{E}\|\hat{\boldsymbol{g}}\|^2 \geq \frac{\mu\phi_C(H_w)}{N} > 0$, and $r_N(\theta) \leq \frac{B\sqrt{2\mathcal{E}(\theta)}}{\sqrt{2B^2\mathcal{E}(\theta) + \mu\phi_C(H_w)/N}} \to 0$.*

Proof 6 *Let the sample-wise stochastic gradient be $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{J}_\theta(\boldsymbol{x})^\top(\boldsymbol{p} - \boldsymbol{y})$. By the law of total covariance,*

$$\Sigma = \mathrm{Cov}(\boldsymbol{g}) = \mathbb{E}_{\boldsymbol{x}}\left[\mathrm{Cov}_{y\sim\boldsymbol{q}}(\boldsymbol{g})\right] + \mathrm{Cov}_{\boldsymbol{x}}(\mathbb{E}_{\boldsymbol{y}}\boldsymbol{g}) \succeq \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{J}^\top(\mathrm{diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\top)\boldsymbol{J}\right]. \tag{28}$$

Let

$$A(\boldsymbol{q}) = \mathrm{diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\top \succeq 0. \tag{29}$$

Since $A(\boldsymbol{q})\mathbf{1} = 0$, its range is contained in $\mathcal{S}_0$, and $\mathrm{tr}A(\boldsymbol{q}) = 1 - \|\boldsymbol{q}\|_2^2$. By Assumption 2,

$$\begin{aligned}
\mathrm{tr}\left(\boldsymbol{J}^\top A(\boldsymbol{q})\boldsymbol{J}\right) &= \mathrm{tr}\left(A(\boldsymbol{q})K_\theta(\boldsymbol{x}, \boldsymbol{x})\right) \\
&= \mathrm{tr}\left(A(\boldsymbol{q})\Pi_0 K_\theta(\boldsymbol{x}, \boldsymbol{x})\Pi_0\right) \\
&\geq \mu\mathrm{tr}A(\boldsymbol{q}) \\
&= \mu(1 - \|\boldsymbol{q}\|_2^2).
\end{aligned} \tag{30}$$

Taking the expectation over x gives

$$\mathrm{tr}\Sigma \geq \mu \mathbb{E}_{\boldsymbol{x}} \left[1 - \|\boldsymbol{q}\|_2^2\right]. \tag{31}$$

We next convert entropy directly into a lower bound on the error mass that holds for any $C \geq 2$. Let

$$\boldsymbol{q}_{\max} = \max_j \boldsymbol{q}_j, P_e = 1 - \boldsymbol{q}_{\max} \in [0, 1 - 1/C]. \tag{32}$$

Since

$$\|\boldsymbol{q}\|_2^2 = \sum_j \boldsymbol{q}_j^2 \leq \boldsymbol{q}_{\max} \sum_j \boldsymbol{q}_j = \boldsymbol{q}_{\max}, \tag{33}$$

we can obtain

$$1 - \|\boldsymbol{q}\|_2^2 \geq P_e. \tag{34}$$

On the other hand, grouping the most probable class against all remaining classes yields

$$H(\boldsymbol{q}) \leq H_b(P_e) + P_e \log(C-1) = F_C(P_e). \tag{35}$$

Its derivatives are

$$F_C'(e) = \log \frac{(C-1)(1-e)}{e} \geq 0, F_C''(e) = -\frac{1}{e(1-e)} < 0. \tag{36}$$

Thus, F_C is monotonically increasing and concave on this interval, and its inverse $\phi_C = F_C^{-1}$ is monotonically increasing and convex. Therefore, $P_e \geq \phi_C(H(q))$, and pointwise, $1 - \|\boldsymbol{q}\|_2^2 \geq \phi_C(H(\boldsymbol{q}))$. Applying Jensen's inequality gives

$$\mathbb{E}_x\left[1 - \|\boldsymbol{q}\|_2^2\right] \geq \mathbb{E}_{\boldsymbol{x}}[\phi_C(H(\boldsymbol{q}))] \geq \phi_C(\mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q})]). \tag{37}$$

Hence,

$$\boxed{\mathrm{tr}\Sigma \geq \mu \phi_C(\mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q})])}. \tag{38}$$

On the other hand,

$$\boldsymbol{g}_{ce} = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{J}^\top(\boldsymbol{p} - \boldsymbol{q})]. \tag{39}$$

Using $\|\boldsymbol{J}\|_2 \leq B$, $\|\boldsymbol{v}\|_2 \leq \|\boldsymbol{v}\|_1$, Pinsker's inequality, and Jensen's inequality,

$$\begin{aligned} \|\boldsymbol{g}_{ce}\| &\leq B\, \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{p} - \boldsymbol{q}\|_2 \\ &\leq B\, \mathbb{E}_{\boldsymbol{x}} \|\boldsymbol{p} - \boldsymbol{q}\|_1 \\ &\leq B\sqrt{2\mathbb{E}_{\boldsymbol{x}}[\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p})]} = B\sqrt{2\mathcal{E}(\theta)}. \end{aligned} \tag{40}$$

To show that the certainty gain vanishes, it suffices to note that, for fixed C, the entropy gradient is uniformly bounded. By Lemma 1, we have

$$\nabla_{\boldsymbol{z}} H = -\boldsymbol{p} \odot (\log \boldsymbol{p} + H\mathbf{1}). \tag{41}$$

Using $|\boldsymbol{u} \log \boldsymbol{u}| \leq e^{-1}$ and $H(\boldsymbol{p}) \leq \log C$, there exists a finite constant M_C depending only on C such that $\|\nabla_{\boldsymbol{z}} H\|_2 \leq M_C$. Therefore,

$$\|\boldsymbol{g}_e\| = \left\|\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{J}^\top \nabla_{\boldsymbol{z}} H]\right\| \leq B M_C. \tag{42}$$

Under gradient flow $\dot{\theta} = -\eta \boldsymbol{g}_{ce}$,

$$|\dot{\mathcal{C}}| = \eta |\langle \boldsymbol{g}_e, \boldsymbol{g}_{ce} \rangle| \leq \eta B^2 M_C \sqrt{2\mathcal{E}(\theta)} \to 0. \tag{43}$$

If the weak modality satisfies

$$\mathbb{E}_{\boldsymbol{x}}[H(\boldsymbol{q}^v)] \geq H_w > 0, \tag{44}$$

then, because ϕ_C is strictly increasing and $\phi_C(0) = 0$, $\phi_C(H_w) > 0$. Hence, $\mathrm{tr}\Sigma \geq \mu \phi_C(H_w)$. For an i.i.d. mini-batch of size N,

$$\mathbb{E}\|\hat{\boldsymbol{g}}\|^2 = \|\boldsymbol{g}_{ce}\|^2 + \frac{1}{N}\mathrm{tr}\Sigma \geq \frac{\mu \phi_C(H_w)}{N} > 0. \tag{45}$$

Finally, by the Cauchy–Schwarz inequality, we obtain

$$r_N(\theta) = \frac{|\langle \boldsymbol{g}_e, \boldsymbol{g}_{ce} \rangle|}{\|\boldsymbol{g}_e\| \sqrt{\mathbb{E}\|\hat{\boldsymbol{g}}\|^2}} \leq \frac{B\sqrt{2\mathcal{E}(\theta)}}{\sqrt{2B^2\mathcal{E}(\theta) + \mu \phi_C(H_w)/N}} \to 0. \tag{46}$$

This completes the proof.

B Analysis of MaxCR

In this section, we analyze the effectiveness and conservativeness of peak regularization. Our analysis aims to provide a local explanation in an idealized setting, while a rigorous treatment that directly accounts for practical DNN training is left to future work.

B.1 Certainty Effect

Proposition 1 (Certainty-Active Component) *Let $m = \arg\max_j \boldsymbol{p}_j$. For $\mathcal{L}_{exc} = \lambda\left(\frac{1}{C} - \boldsymbol{p}_m\right)$, the induced sample-wise rate of certainty change is $\dot{\mathcal{C}}_{exc} = \lambda \boldsymbol{p}_m \Psi(\boldsymbol{p}, m) \geq \lambda \boldsymbol{p}_m^2 (1 - \boldsymbol{p}_m)\boldsymbol{a}_m \geq 0$, and is strictly positive when $\mathbb{C}(\boldsymbol{p}) = \{m\}$. The loss $\mathcal{L}_{sup}$ induces an equal-magnitude change with the opposite sign. This result is independent of the label $\boldsymbol{y}$ and posterior $\boldsymbol{q}$.*

Proof 7 *Since $\nabla_{\boldsymbol{z}} \boldsymbol{p}_m = \boldsymbol{p}_m(\boldsymbol{e}_m - \boldsymbol{p})$, gradient descent gives*

$$\dot{\boldsymbol{z}} = -\nabla_{\boldsymbol{z}} \mathcal{L}_{exc} = \lambda \boldsymbol{p}_m (\boldsymbol{e}_m - \boldsymbol{p}), \tag{47}$$

where $\boldsymbol{e}_m$ is a one-hot vector. Therefore,

$$\begin{aligned} \dot{\mathcal{C}}_{exc} &= -\langle \nabla_{\boldsymbol{z}} H, \dot{\boldsymbol{z}} \rangle \\ &= \lambda \boldsymbol{p}_m \langle \boldsymbol{p} \odot \boldsymbol{a}, \boldsymbol{e}_m - \boldsymbol{p} \rangle \\ &= \lambda \boldsymbol{p}_m (\boldsymbol{p}_m \boldsymbol{a}_m - \Gamma) \\ &= \lambda \boldsymbol{p}_m \Psi(\boldsymbol{p}, m). \end{aligned} \tag{48}$$

By Lemma 3,

$$\boldsymbol{p}_m = \boldsymbol{p}_{\max} \geq e^{-H} \Rightarrow \boldsymbol{a}_m \geq 0. \tag{49}$$

If $\mathbb{C}(\boldsymbol{p}) = \{m\}$, then by Theorem 2,

$$\Psi(\boldsymbol{p}, m) \geq \boldsymbol{p}_m (1 - \boldsymbol{p}_m) \boldsymbol{a}_m > 0. \tag{50}$$

The same argument for $\mathcal{L}_{sup}$ yields the opposite sign.

This completes the proof.

B.2 Comparison with Entropy and Conservativeness

Difference from Entropy Minimization: For $\mathcal{L}_{EM} = \lambda H(\boldsymbol{p})$, we have $\dot{\mathcal{C}}_{EM} = \lambda \|\nabla_{\boldsymbol{z}} H\|^2 = \lambda \sum_j \boldsymbol{p}_j^2 \boldsymbol{a}_j^2$. Near a uniform prediction, $\boldsymbol{a} \to 0$, and therefore $\dot{\mathcal{C}}_{EM} = O(\|\boldsymbol{a}\|^2)$ is second-order small. In contrast, for $\dot{\mathcal{C}}_{exc} = \lambda \boldsymbol{p}_m \Psi(\boldsymbol{p}, m)$, Ψ is first-order in the margin.

Conservativeness: When the top-1 class $m \neq y$, the excitation term reinforces an incorrect prediction. Since $\lambda \leq 1$, $\boldsymbol{p}_y \leq \boldsymbol{p}_m$, $\boldsymbol{p}_y \leq 1 - \boldsymbol{p}_m$, and $\boldsymbol{p}_m^2 + \boldsymbol{p}_y^2 \leq \|\boldsymbol{p}\|_2^2 \leq \boldsymbol{p}_m$, we have

$$\begin{aligned} \frac{d}{dt} \log \frac{\boldsymbol{p}_m}{\boldsymbol{p}_y} &= -(1 + \boldsymbol{p}_m - \boldsymbol{p}_y) + \lambda \boldsymbol{p}_m (1 - \boldsymbol{p}_m + \boldsymbol{p}_y) \\ &\leq -(1 + \boldsymbol{p}_m - \boldsymbol{p}_y) + \boldsymbol{p}_m (1 - \boldsymbol{p}_m + \boldsymbol{p}_y) \\ &\leq \boldsymbol{p}_y - 1 \\ &\leq -\boldsymbol{p}_m \\ &\leq -1/C. \end{aligned} \tag{51}$$

Therefore, we can obtain a uniform negative upper bound.

Consider one step of gradient descent:

$$\begin{aligned} \dot{\boldsymbol{p}}_y &= \boldsymbol{p}_y (1 - \boldsymbol{p}_y - \lambda \boldsymbol{p}_m \boldsymbol{p}_y - \langle \boldsymbol{p}, (\boldsymbol{e}_y - \boldsymbol{p}) + \lambda \boldsymbol{p}_m (\boldsymbol{e}_m - \boldsymbol{p}) \rangle) \\ &= \boldsymbol{p}_y (1 - \boldsymbol{p}_y - \lambda \boldsymbol{p}_m \boldsymbol{p}_y - (\boldsymbol{p}_y - \|\boldsymbol{p}\|_2^2) + \lambda \boldsymbol{p}_m (p_m - \|\boldsymbol{p}\|_2^2)). \end{aligned} \tag{52}$$

Therefore, we have

$$\frac{\dot{\boldsymbol{p}}_y}{\boldsymbol{p}_y} = (1 - 2\boldsymbol{p}_y + \|\boldsymbol{p}\|_2^2) - \lambda \boldsymbol{p}_m (\boldsymbol{p}_y + \boldsymbol{p}_m - \|\boldsymbol{p}\|_2^2). \tag{53}$$

Table I: Notation Definition

Notation	Description
$\mathcal{D}$	Training dataset.
$\boldsymbol{x}_i^a/\boldsymbol{x}_i^v$	Audio/video data point.
$\boldsymbol{y}_i \in \{0,1\}^C$	Category label of i-th data.
Superscript r	r-modality.
$\boldsymbol{x}^r/\boldsymbol{y}$	Data and category labels.
M	The number of data points.
C	The number of category labels.
$f^r(\cdot)$	Encoder of r modality.
θ^r	Parameters of $f^r(\cdot)$.
$\boldsymbol{z}$	Logits of data.
$\boldsymbol{h}^r$	Feature of $\boldsymbol{x}^r$.
$\psi^r(\cdot)$	Classifier of r modality.
ϕ^r	Parameters of $\psi^r(\cdot)$.
$\boldsymbol{p}^r$	Prediction of $\boldsymbol{x}^r$.
$\Phi^r = \{\theta^r, \phi^r\}$	Parameters set.
$\ell(\cdot)$	Cross entropy loss.
$\mathcal{L}_{CE}(\cdot)$	Cross entropy loss over training.
s	Sparsity metric.
τ	Scaling factor.
S	Sparsity score.
$\boldsymbol{B}$	Mini-batch data.
N	Mini-batch size.
σ	Dead zone.
λ	Modality discrepancy ratio.
$\mathcal{L}_{sup}(\cdot)$	Max suppression loss.
$\mathcal{L}_{exc}(\cdot)$	Max excitation loss.
$H(\cdot)$	Entropy loss.
$\boldsymbol{g}_{ce}$	Task gradient.
$\boldsymbol{g}_e$	Entropy gradient.
$\boldsymbol{J}$	Jacobian.
$\mathcal{E}$	Excess risk.
$r_N(\theta)$	Certainty-transfer ratio.

Since $\|\boldsymbol{p}\|_2^2 \leq \boldsymbol{p}_m$, we have $\boldsymbol{p}_m(\boldsymbol{p}_y + \boldsymbol{p}_m - \|\boldsymbol{p}\|_2^2 \geq \boldsymbol{p}_m\boldsymbol{p}_y \geq 0$. So, when $\lambda \leq 1$, we have

$$\begin{aligned}\frac{\dot{\boldsymbol{p}}_y}{\boldsymbol{p}_y} &\geq (1 - 2\boldsymbol{p}_y + \|\boldsymbol{p}\|_2^2) - \boldsymbol{p}_m(\boldsymbol{p}_y + \boldsymbol{p}_m - \|\boldsymbol{p}\|_2^2) \\ &\geq (1 - \boldsymbol{p}_y)[(1 - \boldsymbol{p}_y) - \boldsymbol{p}_m\boldsymbol{p}_y] + \boldsymbol{p}_m^3 \\ &\geq (1 - \boldsymbol{p}_y)^3 + \boldsymbol{p}_m^3 \\ &> 0.\end{aligned} \tag{54}$$

That is, the odds of the incorrect top-1 class against the ground-truth class contract at a rate bounded away from zero uniformly over samples, and the ground-truth class probability strictly increases.

C Notation Definition

We summarize the notation definition in Table I.

D EXPERIMENT DETAILS

D.1 DATASET DETAILS

CREMAD: CREMAD Cao et al. (2014) is an audio-visual dataset designed for speech emotion recognition. It comprises 7,442 video clips of 2∼3 seconds from 91 actors speaking several short words. This dataset includes the six most common emotions.

KSounds: KSounds Arandjelovic & Zisserman (2017) dataset is a commonly used dataset containing 31 action categories that can be recognized visually and auditorily, which contains 19,000 10-second video clips from YouTube.

VGGSound: VGGSound Chen et al. (2020) dataset is an audio-visual dataset in the wild, with nearly 200K 10-second video clips. Each sound-emitting object is also visible in the corresponding video clip in this dataset. After filtering out unavailable videos, 168,618 videos were selected for training and validation, and 13,954 videos for testing in experimental settings.

Twitter: Twitter Yu & Jiang (2019) dataset is a dataset designed for emotion recognition. This dataset consists of tweets collected from Twitter with three different labels: positive, negative, and neutral.

Sarcasm: Sarcasm Cai et al. (2019) dataset is specifically designed for sarcasm detection tasks, containing a large amount of text data labeled as sarcastic or non-sarcastic, which typically draws from various sources, such as social media, news comments, and conversation data, with each entry clearly labeled.

NVGesture: NVGesture Molchanov et al. (2016) dataset is a multimodal dataset specifically designed for gesture recognition, containing three types of data modalities, i.e., RGB, optical flow (OF), and Depth. This dataset includes 25 different gesture categories, covering a variety of common gestures.

D.2 MORE IMPLEMENTATION DETAILS

We report implementation details such as the learning rate and optimizer in the main paper, and further provide additional details here. The batch size is set to be 64 for CREMAD dataset, 32 for KSounds, Twitter, and Sarcasm datasets, 16 for VGGSound dataset, and 2 for NVGesture dataset in training. Following Jiang et al. (2025b), we conduct experiments with three random seeds and report the mean and standard deviation. In addition, following Zhang et al. (2024); Hua et al. (2024); Yang et al. (2024); Jiang et al. (2025b), we obtain multimodal predictions by averaging the predictions from all modalities.

D.3 EXTENSION TO MORE THAN TWO MODALITIES

Although we present our method in the paper using a bimodal setting for simplicity, MaxCR can be readily generalized to more modal scenarios with a simple strategy. Specifically, when computing the modality discrepancy ratio λ, we compare the sparsity score S_t^r of the current modality against the average score $\bar{S}_t$ over all modalities:

$$\lambda = \left| S_t^r - \bar{S}_t \right|. \tag{55}$$

The current modality is then suppressed or excited according to this ratio, which naturally extends MaxCR from the bimodal case to more general multimodal settings.

E ADDITIONAL EXPERIMENTS

E.1 FURTHER VALIDATION EXPERIMENTS

Experimental Details: Unlike naive MML, which is trained solely with concatenation, we additionally introduce unimodal losses into multimodal learning as a baseline, following Wei & Hu (2024); Huang et al. (2025). This uniform learning enhances unimodal gradient signals and partially alleviates modality imbalance. Moreover, in the validation experiments, introducing unimodal losses facilitates the investigation of the relationship between unimodal task loss and the entropy reduction direction in multimodal learning.

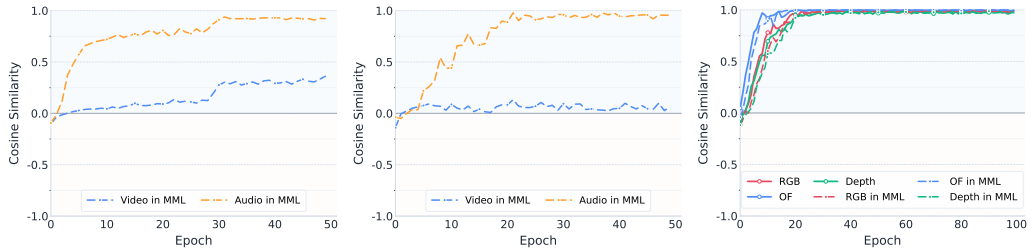


Figure I: Additional experiment between the task gradient and the entropy-reduced gradient on Sarcasm (left), Twitter (middle) and NVGesture (right) datasets.

Table II: Performance with different monitoring metrics on CREMAD dataset.

Method	Audio	Video	Multimodal
Confidence	0.6788	0.6613	0.8427
Entropy	0.6815	0.6694	0.8468
Sparsity (Our)	**0.6855**	**0.6828**	**0.8531**

Table III: Performance of MaxCR with OGM on KSounds dataset.

Method	Audio	Video	Multimodal
OGM	0.5013	0.3165	0.6582
MaxCR	0.5570	**0.5605**	0.7522
MaxCR+OGM	**0.5651**	0.5593	**0.7530**

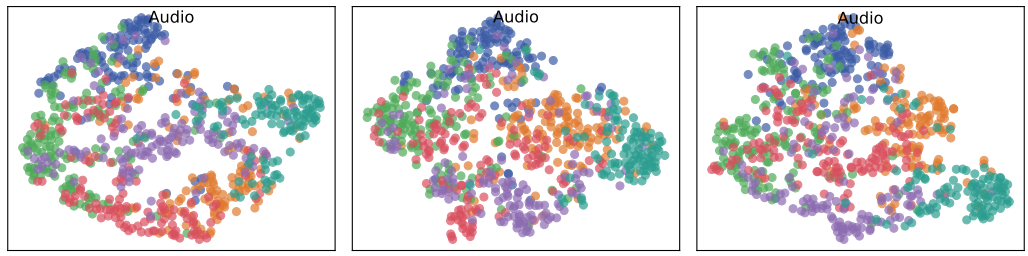
More Results and Analysis: We conduct experiments on image-text datasets, namely Sarcasm and Twitter. As shown in Figure I, similar to the audio-video datasets, the image-text datasets also exhibit $r_N(\theta) \to 0$. A natural question is whether the optimization-certainty decoupling also exists in relatively homogeneous multimodal data, where modality imbalance is less pronounced and the unimodal confidence gap is less evident. To investigate this issue, we further conduct experiments on NVGesture dataset, which consists of three visual modalities, i.e., RGB, optical flow (OF), and Depth. Compared with audio-video settings, these modalities are more homogeneous and exhibit weaker modality discrepancy. As shown in Figure I, the optimization-certainty decoupling is not consistently observed throughout training on NVGesture dataset, and the discrepancy mainly appears in the early stage. Nevertheless, MaxCR still improves the performance. This gain may arise because MaxCR does not directly reshape the optimization dynamics, but instead corrects asymmetric confidence behavior. The result indicates that our method remains beneficial beyond the assumption in the analysis, and also generalizes to multimodal settings where modality imbalance is relatively mild.

E.2 Selection of Monitoring Metric on CREMAD dataset

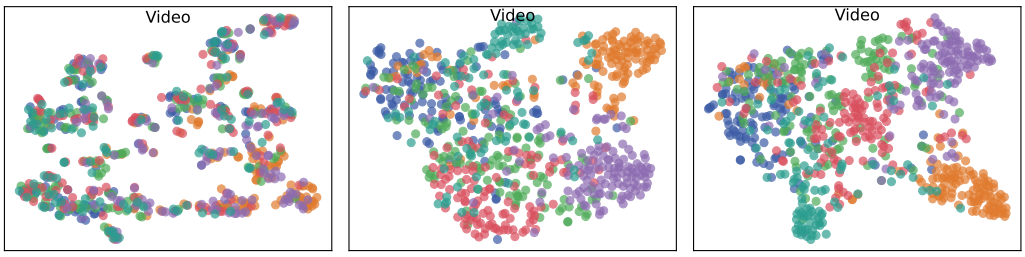
Unlike conventional moving average smoothing that emphasizes historical information, multimodal rebalancing learning requires greater attention to the real-time learning state. Our smoothing of monitoring indicators aims to mitigate state instability in the early training stage. Therefore, the smoothing of the sparsity score progressively reduces the influence of historical information. Furthermore, to evaluate the effectiveness of the sparsity metric, we conduct experiments on CREMAD dataset. As shown in Table II, the sparsity metric achieves the best performance.

E.3 Incorporating Optimization Dynamics

MaxCR focuses on correcting the asymmetric semantic confidence between strong and weak modalities, whereas methods such as OGM aim to rebalance optimization dynamics by modulating their gradient signals. These two mechanisms operate at different levels and are therefore compatible. To verify this compatibility, we further combine MaxCR with OGM, where OGM is used to regulate the gradient updates, and MaxCR is used to regularize their prediction confidence. This combined design allows multimodal learning to benefit from both balanced optimization dynamics and improved modality semantic confidence. We conduct experiments on KSounds dataset. As shown in Table III, introducing OGM for gradient modulation can further improve model performance.



(a). Naive Audio. (b). LFM Audio. (c). MaxCR Audio.



(d). Naive Video. (e). LFM Video. (f). MaxCR Video.

Figure II: The representations of audio and video modalities on CREMAD dataset by t-SNE Van der Maaten & Hinton (2008) across different methods are shown.

E.4 Analysis on Semantic Confidence

To assess MaxCR's impact on semantic confidence, we evaluate the mean top-1 predictive confidence and average number of candidate classes (classes with probabilities exceeding the mean). Assuming C total classes, we experiment on the CREMAD dataset ($C = 6$). As reported in Table IV, the results demonstrate that MaxCR enhances confidence and reduces candidate classes for video modality, while moderately suppressing confidence for audio modality.

Table IV: Comparison of Acc, Top-1 confidence and candidate classes ($> 1/C$).

Method	ACC	Top-1		$> 1/C$	
		Audio	Video	Audio	Video
Naive	0.6361	**0.9234**	0.6643	1.1694	1.6640
OGM	0.6612	0.9232	0.6972	1.1788	1.5121
MaxCR	**0.8531**	0.8768	**0.7783**	1.2742	1.4892

E.5 Visualization Results

In Figure II, we further analyze the proposed MaxCR via t-SNE Van der Maaten & Hinton (2008) feature visualization. Specifically, we compare the visualization results of naive multimodal learning (Naive), the label intervention method LFM, and the proposed method MaxCR. It can be observed that the visual features learned by Naive MML do not exhibit clear decision boundaries, which is consistent with its inferior performance. In contrast, LFM and MaxCR produce higher quality representations. Moreover, the video modality features learned by MaxCR form clearer clusters.