-Supplementary Materials-
Multimodal Learning on Low-Quality Data with Conformal Predictive Self-Calibration

Xun Jiang[1,2], Yufan Gu[2], Disen Hu[2], Yuqing Hou[3], Yazhou Yao[4], Fumin Shen[2],
Heng Tao Shen[1], Xing Xu[1*]
[1]School of Computer Science and Technology, Tongji University
[2]School of Computer Science and Engineering,
University of Electronic Science and Technology of China, [3]Independent Researcher
[4]School of Computer Science and Engineering, Nanjing University of Science and Technology

1. An Overview of Supplementary Materials

In this supplementary materials, we present more details and experimental results which are not listed in the formal paper due to page limit. Specifically, we would present the detailed theoretical proofs, comprehensive implementation details, extended experimental results, and additional analyses that complement the primary text. Overall, this supplementary material is structured as follows:

- **Section 2: Theoretical Proofs.** In this section, we provide the detailed mathematical derivations and proofs for the propositions presented in the main paper, specifically: Proposition 1 for robust feature reconstruction and Proposition 2 for gradient variance reduction.
- **Section 3: Experimental Details.** We provide comprehensive information on our experimental setup, including detailed statistics for all six benchmark datasets, backbone architectures used for each modality, and hyperparameter configurations applied during training.
- **Section 4: Additional Experiments.** we present additional quantitative analysis to further demonstrate the effectiveness of CPSC, including model convergence analysis, sensitivity analysis for key hyperparameters, generalization capability, and visualizations.

Additionally, we also provide our implementation code, which will be publicly available after paper acceptance. More details can be found in our code repository.

2. Theoretical Proofs

We provide the comprehensive mathematical proofs for the key theoretical propositions introduced in the main paper, which underpin the effectiveness of our CPSC framework. Specifically, we present the formal derivations for the following two propositions: (1) **Proposition 1**: Demonstrating the theoretical bound on the expected deviation between the feature reconstructed by the RSC module and an ideal robust representation. (2) **Proposition 2**: Establishing the condition under which the GSC module reduces the effective variance of the stochastic gradient estimate.

Proposition 1 *The expected deviation between the calibrated representation $\tilde{h}^m$ and an ideal robust representation h_*^m is bounded by:*

$$\mathbb{E}[\|\tilde{h}^m - h_*^m\|_2] \leq \frac{1}{K} \sum_{k \in \mathcal{S}^m} \mathbb{E}[\|c_k^m - h_*^m\|_2], \quad (1)$$

where the selection of components into $\mathcal{S}^m$ ensures that $\mathbb{E}[\|c_k^m - h_*^m\|_2]$ is minimized for $k \in \mathcal{S}^m$.

Proof 1 *Following the calculation of the calibrated features in RSC module, we have:*

$$\tilde{h}^m = \frac{1}{K} \sum_{k \in \mathcal{S}^m} c_k^m, \quad (2)$$

where c_k^m denotes a component and $\mathcal{S}^m$ is the component set selected by RSC module. Let h_*^m be the ideal robust representation. The reconstruction error is expressed as:

$$\begin{aligned} \|\tilde{h}^m - h_*^m\|_2 &= \left\| \frac{1}{K} \sum_{k \in \mathcal{S}^m} c_k^m - h_*^m \right\|_2 \\ &= \frac{1}{K} \left\| \sum_{k \in \mathcal{S}^m} (c_k^m - h_*^m) \right\|_2 \\ &\leq \frac{1}{K} \sum_{k \in \mathcal{S}^m} \|c_k^m - h_*^m\|_2. \end{aligned}$$

*Corresponding author.

Taking expectation on both sides:

$$\mathbb{E}[\|\tilde{h}^m - h_*^m\|_2] \leq \frac{1}{K} \sum_{k \in \mathcal{S}^m} \mathbb{E}[\|c_k^m - h_*^m\|_2]. \quad (3)$$

Here we give the reliability scoring process as follows, which has been formualted in our formal paper:

$$r_k^m = 1 - \frac{rank\,[y, C(c_k^m)]}{|C(c_k^m)|}. \quad (4)$$

The reliability scoring mechanism above ensures that components with higher scores are more likely to produce correct predictions, which correlates with being closer to the ideal representation h_*^m in the feature space. Therefore, by selecting the top-K components with highest reliability scores, we minimize the expected deviation $\mathbb{E}[\|c_k^m - h_*^m\|_2]$ for $k \in \mathcal{S}^m$.

Proposition 2 *For a convex loss function $\mathcal{L}$, the Gradient Self-Calibration module with linear weighting reduces the effective variance of the stochastic gradient estimate when $w(\rho_i)$ is positively correlated with $\|\nabla \mathcal{L}_i\|_2$.*

Proof 2 *Let $\mathcal{L}(\theta)$ be the overall expected loss. The true gradient is $\nabla \mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}}[\nabla \mathcal{L}(\mathbf{x}, y)]$. The variance of the standard SGD estimator is:*

$$Var(G_{std}) = Var\left(\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} \nabla \mathcal{L}_i\right) = \frac{1}{|\mathcal{B}|} Var(\nabla \mathcal{L}_i).$$

According to our definition in our formal paper, the GSC estimator G_{GSC} uses a per-sample weighting $w(\rho_i)$. To ensure the estimator is unbiased, we normalize the weights by their batch average $\bar{w} = \frac{1}{|\mathcal{B}|}\sum_{i \in \mathcal{B}} w(\rho_i)$, approximating $\mathbb{E}[w(\rho)]$ in practice.

$$G_{GSC} = \frac{1}{|\mathcal{B}| \cdot \bar{w}} \sum_{i \in \mathcal{B}} w(\rho_i) \nabla \mathcal{L}_i,$$

then the variance of the GSC estimator is:

$$\begin{aligned} Var(G_{GSC}) &= \frac{1}{|\mathcal{B}|} Var\left(\frac{w(\rho_i)}{\bar{w}} \nabla \mathcal{L}_i\right) \\ &= \frac{1}{|\mathcal{B}|\bar{w}^2} Var(w(\rho_i) \nabla \mathcal{L}_i). \end{aligned}$$

Following $Var(X) = \mathbb{E}[\|X\|^2] - \|\mathbb{E}[X]\|^2$, we have:

$$\begin{aligned} &Var(G_{std}) - Var(G_{GSC}) = \\ &\frac{1}{|\mathcal{B}|}\left[\left(\mathbb{E}[\|\nabla \mathcal{L}_i\|^2] - \frac{1}{\bar{w}^2}\mathbb{E}[\|w(\rho_i)\nabla \mathcal{L}_i\|^2]\right)\right. \\ &\left. + \left(\frac{1}{\bar{w}^2}\|\mathbb{E}[w(\rho_i)\nabla \mathcal{L}_i]\|^2 - \|\mathbb{E}[\nabla \mathcal{L}_i]\|^2\right)\right]. \end{aligned}$$

Since we assume the GSC estimator is unbiased, we have $\frac{1}{\bar{w}}\mathbb{E}[w(\rho_i)\nabla \mathcal{L}_i] = \mathbb{E}[\nabla \mathcal{L}_i]$. Therefore, the second parenthetical term is approximately zero:

$$\mathbb{E}[\|\nabla \mathcal{L}_i\|^2] \geq \mathbb{E}\left[\left(\frac{w(\rho_i)}{\bar{w}}\right)^2 \|\nabla \mathcal{L}_i\|^2\right]$$

According to our assumption that high reliability ρ_i correctly identifies samples whose large gradients are informative and should be amplified, whereas low reliability correctly identifies samples whose gradients are corrupted and should be suppressed. Our selective up-weighting of trustworthy components and down-weighting of untrustworthy components effectively reduces the impact of noise on the overall gradient direction. By defining $w(\rho)$ to be positively correlated with the magnitude of the informative gradient, the GSC mechanism effectively separates the reliable signal from the noise, leading to a tighter distribution of the stochastic gradient around the true gradient $\nabla \mathcal{L}(\theta)$. This results in a stable optimization, validating the statement $Var(G_{GSC}) \leq Var(G_{std})$.

3. Experimental Details

3.1. Datasets and Evaluation Metrics

We evaluate our CPSC method on six multimodal datasets, three imbalanced multimodal datasets: CREMA-D [1], AVE [13], Kinetics Sounds [7], and three noisy multimodal datasets: SUN RGB-D [12], NYU Depth V2 [11] and MVSA-Single [9]. We summarize the details of these datasets as follows:

- **CREMA-D** [1]: The CREMA-D dataset contains 7,442 video clips. It is an audio-visual dataset designed for emotion recognition research. It contains over 7,000 clips of 91 actors expressing six basic emotions, including anger, disgust, fear, happy, neutral, and sad. In the experiment, we adapted audio and video as our input modality. And we use 6,698 samples as our training set, and the rest of the samples are used as test set.
- **Kinetics Sounds** [7]: The Kinetic Sounds dataset is a multimodal audio-visual benchmark tailored for tasks like sound recognition and generation. It contains over 1,000 short video clips capturing humans performing actions such as hitting, scraping, and tapping various objects, with precisely synchronized audio and visual modalities. Following the setup in prior work [16], we focus on 31 selected action-sound categories for multimodal learning. The dataset is split into 20,383 clips for training and 1,337 clips for testing.
- **AVE** [13]: The AVE dataset consists of 4,143 10-second videos. It is designed for audio-visual event localization. It contains 4,143 video clips, each 10 seconds long, covering 28 event categories where both audio and visual components are present. Among all the samples, we use 3,339

Table 1. Comparisons with recent state-of-the-art methods under the Imbalanced Multimodal Learning settings. Note that $Acc_{\{m,\,a,\,v\}}$ denote multimodal, audio, and visual classification performance.

Method	Kinetics Sounds				CREMA-D				AVE			
	Acc_m	Acc_a	Acc_v	Avg	Acc_m	Acc_a	Acc_v	Avg	Acc_m	Acc_a	Acc_v	Avg
OGM-GE (CVPR'22) [10]	66.79	51.09	37.86	51.91	71.14	61.29	39.27	57.23	69.12	62.45	27.39	52.99
Greedy (ICML'22) [17]	65.32	50.58	35.97	50.62	69.31	62.49	38.23	56.68	69.66	60.76	38.70	56.37
PMR (CVPR'23) [2]	65.70	52.47	34.52	50.90	75.54	63.04	71.24	69.27	70.89	63.18	35.57	56.55
AGM (ICCV'23) [8]	66.17	51.31	34.83	50.77	77.86	63.34	37.54	59.58	71.04	62.44	40.96	58.15
ReconBoost (ICML'24) [5]	70.85	56.23	50.27	59.12	79.82	60.23	73.01	71.02	71.35	61.20	39.06	57.20
MMPareto (ICML'24) [16]	70.13	56.40	53.05	59.86	78.53	67.38	70.26	72.06	75.81	64.34	45.39	61.85
LFM (NeurIPS'24) [4]	72.53	57.98	56.43	62.31	86.02	66.53	75.27	75.94	68.58	64.35	44.89	59.27
InfoReg (CVPR'25) [6]	72.00	57.21	53.57	60.93	76.28	64.19	70.62	70.36	74.19	63.78	42.54	60.17
IPRM (IJCAI'25) [20]	74.82	59.76	58.34	64.31	85.35	65.28	76.41	75.68	74.61	65.11	43.89	61.20
ARL (ICCV'25) [14]	74.38	58.26	59.74	64.12	83.79	65.92	72.18	73.96	70.42	63.78	39.61	57.94
DGL (ICCV'25) [15]	74.78	52.89	60.11	62.59	82.52	65.36	74.84	74.24	73.89	64.30	42.16	60.11
CPSC (Ours)	**76.08**	**61.54**	**61.83**	**66.48**	**87.83**	**67.74**	**80.38**	**78.65**	**77.66**	**66.93**	**45.65**	**63.41**

samples as our training set, and the rest of the samples are used as testing set.

- **MVSA-Single** [9]: The MVSA-Single dataset is a widely used benchmark for multimodal sentiment analysis, including positive, negative, and neutral. The dataset consists of 4,869 samples that collected from social media, and the official data split includes 3,899 training samples, 490 validation samples, and 480 test samples, facilitating standard experimental comparisons.
- **NYU Depth V2** [11]: This dataset is a widely used benchmark in indoor scene classification. Following previous work [22], we reorganize the 27 categories into 10 categories with 9 usual scenes and one "others" category. we divide the 1,449 samples into a training set and a test set with 795 and 654 samples respectively.
- **SUN RGB-D** [12]: This dataset is designed for indoor scene understanding using RGB and depth modalities. It contains 10,335 RGBD images providing diverse viewpoints and scene types. Following previous work [22], we use the 19 major scene categories of SUN RGB-D, each of which contains at least 80 images.

As for the metrics, we evaluate the model under two settings. For the imbalanced settings [16], we report the multimodal and unimodal accuracy on CREMA-D, AVE, and Kinetics-Sounds to assess the balanced use of modalities. For the robustness settings [22], we report the multimodal accuracy on SUN RGB-D, NYU Depth V2, and MVSA under synthetic noise corruptions like Gaussian and Salt-Pepper noise.

3.2. Implementation Details

In our experiments, we followed the former researches [4, 22] on imbalanced multimodal learning and robust multimodal learning to process both visual and acoustic modalities. Concretely, for audio-visual datasets, the acoustic modality is transformed into 257 $\times$ 1,004 spectrograms, and we randomly sample multiple frames from 10-frame video clips. The balance factors λ_1 and λ_2 are set to 0.8 and 0.2, while the hyperparameters a and b that control the intensity and baseline of the gradient calibration are set to 0.6 and 0.6, respectively. For the RGB-Depth datasets SUN RGB-D and NYU Depth V2, we similarly adopted ResNet18 as the backbone to extract features from both the RGB image and the depth image. For the text-image dataset MVSA, we employed a more powerful architecture to handle the heterogeneous modalities: ResNet152 was utilized for image feature extraction, while the textual content was encoded using a pre-trained BERT model to obtain sentence-level representations. We adopt SGD as the optimizer, with momentum of 0.9 and weight decay parameter of 1e-1, where the initial learning rate is set to 1e-2 with batch size of 64. And for the RGB-Depth datasets and text-image dataset, we implemented the Adam optimizer for both datasets, with learning rates of 1e-4 and 5e-5, respectively.

4. Additional Experiments

In this section, we present supplementary experiments that further validate the effectiveness and robustness of the proposed CPSC framework across multiple datasets and settings. Specifically, we provide (1) More comparative results with more baseline method on the six benchmark datasets. (2) Convergence and ablation studies to quantify the contributions of the RSC and GSC modules, (3) Sensitivity analyses for key hyperparameters, (4) Transferability tests where CPSC is plugged into different baseline methods, and (5) Qualitative visualizations that illustrate improvements in feature separability.

4.1. Additional Comparative Results

To provide a more thorough assessment of our proposed CPSC framework, we conduct an extended comparative study against several well-established methods from the

Table 2. Comparative results under the Robust Multimodal Learning settings. Note that ϵ is the noise strength.

MVSA-Single					
Method	Clean	Gaussian@ϵ		Salt-Pepper@ϵ	
		5.0	10.0	5.0	10.0
QMF (ICML'23) [22]	78.07	73.85	61.28	73.90	60.41
EAU (CVPR'24) [3]	79.15	73.34	61.78	73.69	60.46
MMPareto (ICML'24)	64.16	52.05	45.15	60.31	54.33
ECML (AAAI'24) [18]	76.83	71.28	61.03	72.13	61.04
CRMT (ICLR'24) [21]	65.33	53.78	53.33	55.62	44.54
NLC (AAAI'25) [19]	73.79	65.39	58.98	66.64	57.28
IPRM (IJCAI'25) [20]	75.84	71.25	60.69	70.13	58.26
ARL (ICCV'25) [14]	75.76	70.89	60.74	70.49	59.82
CPSC (Ours)	**80.07**	**74.12**	**63.32**	**73.95**	**61.27**
NYU Depth V2					
Method	Clean	Gaussian@ϵ		Salt-Pepper@ϵ	
		5.0	10.0	5.0	10.0
QMF (ICML'23) [22]	70.09	61.62	55.60	58.50	45.69
ECML (AAAI'24) [18]	71.72	62.08	54.58	57.57	44.93
EAU (CVPR'24) [3]	72.05	62.54	56.23	58.44	46.21
MMPareto (ICML'24)	71.67	60.55	53.32	55.81	44.18
CRMT (ICLR'24) [21]	66.80	55.93	45.43	54.66	43.12
NLC (AAAI'25) [19]	67.33	54.90	45.04	56.02	44.66
IPRM (IJCAI'25) [20]	70.13	58.16	52.69	56.82	43.37
ARL (ICCV'25) [14]	68.72	56.93	49.81	57.13	44.25
CPSC (Ours)	**73.12**	**64.15**	**57.32**	**61.22**	**47.40**
SUN RGB-D					
Method	Clean	Gaussian@ϵ		Salt-Pepper@ϵ	
		5.0	10.0	5.0	10.0
QMF (ICML'23) [22]	61.98	53.40	48.58	52.49	40.53
ECML (AAAI'24) [18]	59.82	52.46	46.71	51.92	39.29
EAU (CVPR'24) [3]	55.68	49.39	44.23	50.38	38.38
MMPareto (ICML'24)	57.89	48.12	43.77	49.90	38.91
CRMT (ICLR'24) [21]	50.32	41.37	35.33	42.18	34.70
NLC (AAAI'25) [19]	52.75	43.57	38.49	45.07	37.25
IPRM (IJCAI'25) [20]	58.67	50.82	45.16	49.98	37.29
ARL (ICCV'25) [14]	56.29	50.78	45.92	50.34	38.46
CPSC (Ours)	**62.12**	**54.11**	**49.10**	**53.37**	**41.28**

multimodal learning literature. This supplements the comparisons presented in our main paper, which focused on recent state-of-the-art methods. In our supplementary materials, we provide the expanded set of results, presented in Table 1 and 2, which contains more counterpart methods [2, 8, 17, 18, 21, 22]. From the comprehensive results, we list the following key observations: (1) Our CPSC method consistently maintains a significant performance advantage over all heuristic and early robust learning baselines across all datasets and evaluation settings. (2) Notably, on datasets with severe implicit imbalance, CPSC's self-calibration mechanism consistently yields gains in multimodal accuracy compared to the strongest classical fusion method, demonstrating the inadequacy of fixed fusion strategies. (3) Furthermore, under high-intensity noise corruption, CPSC exhibits superior stability, underscoring the efficacy of our RSC module in filtering corrupted feature components compared to non-uncertainty-aware robust methods. These extended comparisons firmly establish the necessity and efficacy of the CPSC framework as a unified solution for low-quality multimodal data.

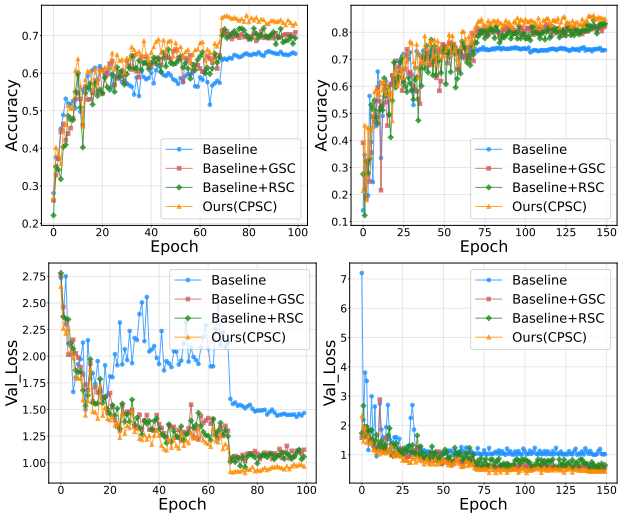

Figure 1. Analysis of model training convergence on CREMA-D and Kinetics Sounds datasets.

4.2. Training Convergence

To evaluate the contributions of the RSC and GSC modules to model performance, we conduct a comprehensive analysis of key performance metrics, specifically accuracy improvement and loss reduction on the validation sets of the CREMA-D and Kinetics Sounds datasets. The experimental design is as follows: under identical training configurations, we compare four model variants: the baseline model, the model with only RSC, the model with only GSC, and the complete model integrating both RSC and GSC modules. The metric progression curves during training are recorded and visualized in Fig. 1. Experimental results demonstrate that introducing either RSC or GSC individually significantly improves accuracy and reduces loss, while the complete model achieves optimal performance with the highest accuracy and lowest loss, validating the effectiveness and synergistic benefits of the proposed modules.

4.3. Analysis on Hyperparameters

We conduct comprehensive ablation studies during model training to optimize the configuration of balancing factors λ_1 and λ_2 in the RSC module for enhancing the robustness and diversity of feature representations. Concretely, we design a systematic hyperparameter search experiment. Based on the consistency constraint and diversity constraint in Eq. (7), we perform a grid search over different combinations of λ_1 and λ_2 values, where λ_1 ranges over [0.6, 0.7, 0.8, 0.9] and λ_2 ranges over [0.1, 0.2, 0.3, 0.4]. For each parameter combination, we train the CPSC model and evaluate its performance on key metrics, thereby analyzing the impact

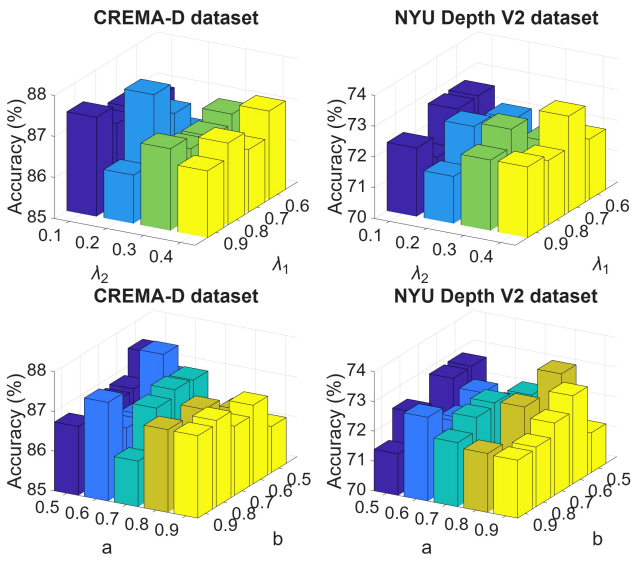

Figure 2. Analysis of hyperparameters in RSC and GSC modules on CREMA-D and NYU Depth V2.

of λ_1 and λ_2 on feature decomposition and fusion. Moreover, we also conduct similar experiments with the same settings on the hyperparameters a and b in our GSC module. The experiment is conducted on two benchmark datasets, CREMA-D and NYU Depth V2, to comprehensively evaluate the parameter performance across different modalities and data qualities.

According to the experimental results in Fig. 2, we observe that the values of λ_1 and λ_2 significantly affect the model performance, where λ_1 controls the consistency between feature components and the original features, while λ_2 regulates the diversity among components. Ultimately, on the CREMA-D dataset, the optimal solution is achieved when λ_1 and λ_2 equal 0.8 and 0.2, respectively; while on the NYU Depth V2 dataset, the optimal values are $\lambda_1 = 0.7$ and $\lambda_2 = 0.4$. As for the a and b, we can observe that the model is more sensitive to a compared to b. This is because a performs as a scale factor for the reliability scores, which is significant for our GSC module.

4.4. Analysis on Prediction Coverage Rate

To further validate the effectiveness of our proposed CP-based strategy, we also conduct additional analysis by quantifying the prediction coverage rate and average set size of audio and video modality. Specifically, as illustrated in Fig. 3, we show how the The prediction coverage rate and average set size fluncates along the training epoch on the CREMA-D dataset. We can observe that the coverage rate diverges to the target level $1 - \alpha$, where the risk factor α is set to 0.2 in our paper. Moreover, the average coverage set size also diverges to a relative size. These results further demonstrate that our proposed CP-based strategy performs well in the overall framework.

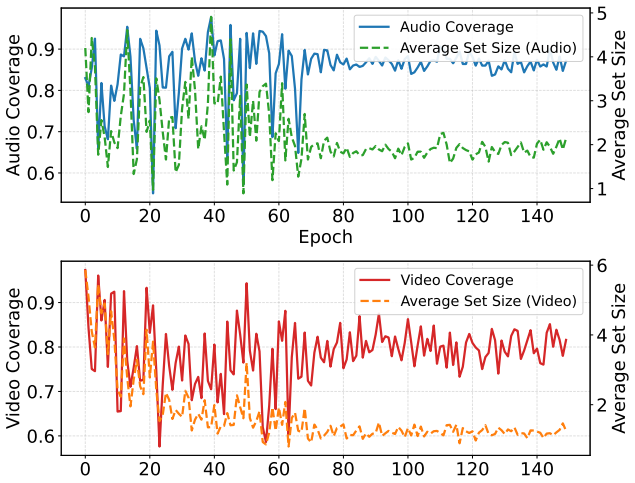

Figure 3. Analysis on the fluncation of prediction coverage rate and average set size on CREMA-D dataset.

Table 3. Model generalization analysis by incorporating key modules into different baseline models on Kinetics Sounds (KS) and CREMA-D (CD) datasets.

Dataset	Optimizer	Acc_m	Acc_a	Acc_v	Avg
CD	LFM	86.02	66.53	75.27	75.94
	LFM+Ours	**87.10**	**67.18**	**77.53**	**77.27**
	ARL	83.79	65.92	73.96	72.18
	ARL+Ours	**84.36**	**66.74**	**75.28**	**75.46**
KS	LFM	72.53	57.98	56.43	62.31
	LFM+Ours	**73.78**	**58.12**	**57.67**	**63.19**
	ARL	74.38	58.26	59.74	64.12
	ARL+Ours	**75.97**	**58.61**	**61.76**	**65.45**

4.5. Analysis on Model Generalization

To validate the transferability and general applicability of our proposed CPSC method, we integrate it as a plug-and-play module into several state-of-the-art multimodal learning frameworks, including LFM [4] and ARL [14]. Specifically, on the CREMA-D and Kinetics Sounds datasets, we select three distinct baseline models, including MMPareto, LFM, and ARL, and conduct comparative experiments with and without the incorporation of the CPSC module. The experimental results are summarized in Table 1. Based on the empirical data, we observe that in the majority of cases, the introduction of our CPSC module leads to consistent improvements in multimodal accuracy (Acc_m), visual accuracy (Acc_v), and average accuracy (Avg). These systematic comparative experiments effectively demonstrate that our CPSC method exhibits strong transferability and general applicability.

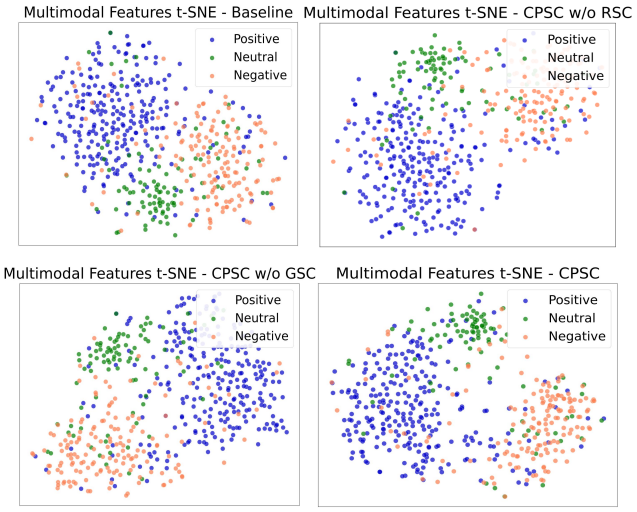

Figure 4. Analysis of feature representations concerning different key modules on the MVSA-Single dataset.

4.6. Additional Qualitative Analysis

To visually validate the enhancement of feature discriminability by the RSC and GSC modules, we conduct a feature space visualization analysis as shown in Fig. 4. Specifically, we train four model variants on the MVSA-Single dataset: the baseline model (Baseline), the model with only GSC (CPSC w/o RSC), the model with only RSC (CPSC w/o GSC), and the complete model (CPSC). We then employ t-SNE dimensionality reduction to project the multimodal features output by each model onto a two-dimensional plane. By comparing the feature distribution maps generated by different models, we qualitatively assess their intra-class compactness and inter-class separation. The experimental results are presented in the aforementioned figure set. From these feature distribution maps, it can be observed that: in the baseline model, the three categories of sample points are mixed with blurred boundaries. After introducing the GSC or RSC module individually, the feature space begins to exhibit preliminary clustering trends, yet considerable overlap remains between different categories. The complete CPSC model achieves the optimal visualization effect, significantly alleviating feature overlap among the three emotional categories.

References

[1] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 2

[2] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. 3, 4

[3] Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26876–26885, 2024. 4

[4] Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. *Advances in Neural Information Processing Systems*, 37: 117363–117387, 2024. 3, 5

[5] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. *arXiv preprint arXiv:2405.09321*, 2024. 3

[6] Chengxiang Huang, Yake Wei, Zequn Yang, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25854–25863, 2025. 3

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[8] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023. 3, 4

[9] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *International conference on multimedia modeling*, pages 15–27. Springer, 2016. 2, 3

[10] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022. 3

[11] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 2, 3

[12] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 2, 3

[13] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 2

[14] Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2250–2259, 2025. 3, 4, 5

[15] Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 22879–22888, 2025. 3

[16] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024. 2, 3

[17] 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 *International Conference on Machine Learning*, pages 24043–24055, 2022. 3, 4

[18] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 16129–16137, 2024. 4

[19] Shilin Xu, Yuan Sun, Xingfeng Li, Siyuan Duan, Zhenwen Ren, Zheng Liu, and Dezhong Peng. Noisy label calibration for multi-view classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21797–21805, 2025. 4

[20] Yang Yang, Xixian Wu, and Qing-Yuan Jiang. Towards equilibrium: An instantaneous probe-and-rebalance multimodal learning approach. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 3552–3560, 2025. 3, 4

[21] Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. *arXiv preprint arXiv:2402.06244*, 2024. 4

[22] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769, 2023. 3, 4