Supplementary Material For Paper: PaSE: Prototype-aligned Calibration and Shapley-based Equilibrium for Multimodal Sentiment Analysis

Anonymous submission

Baseline Models

In this paper, we select a diverse set of current state-of-the-art baselines to conduct a comprehensive comparison, which includes:

- MuLT (Tsai et al. 2019): proposes a Multimodal Transformer with directional crossmodal attention to model unaligned multimodal sequences by adaptively attending to interactions across modalities and time steps without explicit alignment.
- MAG-BERT (Rahman et al. 2020): proposes the Multimodal Adaptation Gate to enhance BERT and XLNet for multimodal language tasks by dynamically adjusting their internal representations based on visual and acoustic inputs during fine-tuning, achieving human-level sentiment analysis performance.
- SelfMM(Yu et al. 2021): proposes a self-supervised framework, the model generates unimodal labels to enable joint training of multimodal and unimodal tasks. A dynamic weight-adjustment strategy prioritizes samples with conflicting modality supervisions, enhancing the capture of nuanced differences.
- HyCon (Mai et al. 2022): employs an adapter-based architecture to inject domain-specific knowledge into general pretrained models, enabling joint learning of modality-specific and universal representations
- ConKI (Yu et al. 2023): employs an adapter-based architecture to inject domain-specific knowledge into general pretrained models, enabling joint learning of modality-specific and universal representations.
- ConFEDE (Yang et al. 2023): introduces contrastive feature decomposition to split text, audio, and visual modalities into similarity and dissimilarity components, guided by text-centric contrastive learning.
- CLGSI (Yang, Dong, and Qiang 2024): introduces a sentiment intensity-guided contrastive learning approach with weighted sample selection and a novel GLFK fusion mechanism for effective multimodal feature extraction.
- MCL-MCF (Fan et al. 2024): proposes a multi-level contrastive learning and multi-layer convolution fusion framework, which progressively mitigates modality heterogeneity through three-level contrastive learning (unimodal, cross-modal, and high-level fusion) and enhances feature fusion via tensor convolution.
- MFON (Zhang, Wei, and Zou 2025): introduces a Modal Feature Optimization Network (MFON) with a Modal Prompt Attention (MPA) mechanism, identifying under-optimized modalities and focusing on their features via task-specific prompts.
- GLoMo (Zhuang et al. 2024): uses modality-specific mixture of experts layers to integrate diverse local representations within each modality and a global-guided fusion module to effectively combine global and local information
- EUAR (Gao et al. 2024): introduces the Enhanced Experts with Uncertainty-Aware Routing (EUAR), integrating the Mixture of Experts approach to dynamically adapt the network via conditional computation, refining experts to capture data uncertainty and using a U-loss-based routing mechanism to direct samples to low-uncertainty experts for noise-free feature extraction in multimodal sentiment analysis.
- KEBR (Zhu et al. 2024): utilizes text-based cross-modal fusion to inject non-verbal information into text representations, applying a multimodal cosine constrained loss to balance joint learning
- Semi-IIN (Lin et al. 2025): proposes a semi-supervised intra-modal and inter-modal interaction learning network for multimodal emotion analysis. Semi-IIN integrates a masked attention mechanism and a gating mechanism, enabling effective dynamic selection after independently capturing intra-modal and inter-modal interaction information.
- MSAmba (He et al. 2025): composes an Intra-modal Sequential Mamba (ISM) module and a Cross-modal Hybrid Mamba (CHM) module to enhance cross-modal interactions.

Implementation Details

To prevent overfitting in the first stage of our dual-phase training, we employ a warm-up period to ensure stable modality-specific learning. When validation performance plateaus, the training automatically transitions to the second phase, activating Shapley-based modality balancing. To ensure fairness, we adopt a feature extraction framework aligned with existing SOTA methods. Specifically, we use Facet, COVAREP (Degottex et al. 2014), and BERT (Devlin et al. 2019) as the

Dataset	# Train	# Validation	# Test
CMU-MOSI	1,284	229	686
CMU-MOSEI	16,326	1,861	4,659
IEMOCAP	2,717	798	938

Table 1: Datasets statistics for fine-tuning and testing

feature extractors for the visual, acoustic, and textual modalities, respectively, on the CMU-MOSI and CMU-MOSEI datasets. We use Adam optimizer with a learning rate of 1e-5, a batch size of 64, and train for 200 epochs. We set $\gamma = 0.98$ for prototype updates, with $\lambda = 0.01$, $\mu = 0.1$, and alignment-enhancing factors $\alpha = 0.1$, $\beta = 0.05$ to improve cross-modal consistency. All experiments are conducted on a single NVIDIA A100 GPU. To mitigate overfitting in the first phase of our proposed dual-phase training strategy, we introduce a warm-up period during the initial training phase, allowing sufficient learning of modality-specific representations. If task performance on the validation set does not significantly improve for several consecutive epochs, the model automatically transitions to the second phase, where a Shapley value-based modality balancing mechanism is activated.

Details of Datasets

PaSE is evaluated on multiple tasks, including multimodal sentiment analysis and multimodal emotion recognition, using three widely adopted benchmark datasets: CMU-MOSI, CMU-MOSEI, and IEMOCAP. As summarized in Table 1, detailed statistics and descriptions of these datasets are provided below.

CMU-MOSI: consists of 2,199 single-person short video clips, with the training set (1,284 samples), validation set (229 samples), and test set (686 samples) configured according to the standard division.

CMU-MOSEI: contains 22,856 movie review videos from YouTube, divided according to the widely accepted academic standard: 16,326 training samples, 1,871 validation samples, and 4,659 test samples. Both datasets use manually annotated labels, with sentiment intensity scores on a continuous scale from -3 to 3, corresponding to five levels of sentiment intensity: "highly negative - negative - neutral - mildly positive - highly positive."

IEMOCAP: consists of 4,453 video clips, including 2,717 training samples, 798 validation samples, and 938 test samples. Following RAVEN (Wang et al. 2019), we adopt four emotion categories (*Happy*, *Sad*, *Angry*, and *Neutral*) for sentiment recognition. For evaluation, we report the F1 score for each category.

More Details of Evaluation Metrics

We report both classification and regression results, averaged over five runs with different random seeds. For classification tasks, we provide multiclass accuracy and F1 scores. Specifically, for the CMU-MOSI and CMU-MOSEI datasets, we

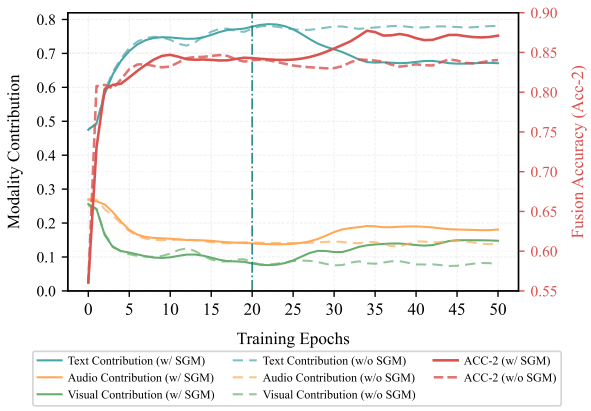

Figure 1: Performance evaluation of modality contribution and fusion ACC-2 on CMU-MOSI. w/o SGM: no modulation is applied throughout training, following conventional methods; w/ SGM: the SGM module is introduced starting from epoch 20.

evaluate both 2-class (Acc-2) and 7-class (Acc-7) accuracy. For IEMOCAP dataset, we provide F1-scores. For regression tasks, we report the Mean Absolute Error (MAE) and the Correlation (Corr). Except for MAE, higher values indicate better performance. For the MOSI and MOSEI datasets, the Acc-2 and F1-score have two forms using the segmentation marker '-/-': the first represents negative/non-negative (including zero) and the second represents negative/positive (excluding zero).

Effectiveness of Dual-Phase Optimization

In most multimodal sentiment analysis (MSA) tasks, the text modality naturally holds an informational advantage, which often leads to the dominant modality suppressing the expressive capacity of other modalities during training. PaSE adopts a dual-phase training strategy: the first phase focuses on constructing clear semantic structures for each modality, while the second phase introduces Shapley-based gradient modulation (SGM) to adjust modality contributions. This design is based on a key observation: introducing Shapley-based modulation too early—before stable modality representations are formed—may lead to unstable optimization.

Figure 1 illustrates the effectiveness of this dual-phase mechanism. The left Y-axis in the figure represents the contributions of each modality, while the red right Y-axis indicates the Acc-2 performance. The red dashed line corresponds to the full training process without SGM, and the red solid line shows the result of introducing SGM after the 20th epoch. As shown, the model nearly converges within the first 20 epochs; after introducing SGM, the contribution of the text modality slightly decreases, but the overall performance improves by approximately 4%. SGM addresses this issue by leveraging Shapley values to measure the marginal contribution of each modality. In the later stages of training, it applies moderate gradient suppression to the dominant modality (text), thereby increasing the participation of weaker modalities. This leads

Model	Training Time	Params
SelfMM	7.7h	121,835,723
GLoMo	9.3h	109,818,887
EUAR	3.9h	110,436,422
KEBR	4.6h	127,467,631
PaSE	4.0h	114,795,953

Table 2: Comparison of training efficiency and parameters across different MSA models.

to more balanced and generalizable fused representations, ultimately improving overall performance.

Efficiency Analysis

To better evaluate the efficiency of our model, we analyze the computational complexity of its key components. PaSE consists of three modules: PCL, with complexity $O(B \times d)$; CAL, with complexity $O(B \times d \times \log d)$, attributed to the use of Entropic Optimal Transport for cross-modal alignment; and SGM, with complexity $O(M \times B)$, where B is the batch size, d is the feature dimension, K is the number of classes, and M is the number of modalities. Overall, the time complexity of PaSE remains linear with respect to input size (i.e., $O(n)$) and is computationally tractable. As shown in Table 2, it is comparable to existing MSA methods such as SelfMM, ConKI, and KEBR, demonstrating that the improved performance of PaSE does not come at the cost of excessive computational overhead.

References

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, 960–964. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Fan, C.; Zhu, K.; Tao, J.; Yi, G.; Xue, J.; and Lv, Z. 2024. Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 16(1): 207–222.

Gao, Z.; Hu, D.; Jiang, X.; Lu, H.; Shen, H. T.; and Xu, X. 2024. Enhanced Experts with Uncertainty-Aware Routing for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9650–9659.

He, X.; Liang, H.; Peng, B.; Xie, W.; Khan, M. H.; Song, S.; and Yu, Z. 2025. MSAmba: Exploring Multimodal Sentiment Analysis with State Space Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1309–1317.

Lin, J.; Wang, Y.; Xu, Y.; and Liu, Q. 2025. Semi-IIN: Semi-supervised Intra-inter modal Interaction Learning Network for Multimodal Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1411–1419.

Mai, S.; Zeng, Y.; Zheng, S.; and Hu, H. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3): 2276–2289.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2020, 2359.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 7216–7223.

Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7617–7630.

Yang, Y.; Dong, X.; and Qiang, Y. 2024. CLGSI: a multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2099–2110.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10790–10797.

Yu, Y.; Zhao, M.; Qi, S.-a.; Sun, F.; Wang, B.; Guo, W.; Wang, X.; Yang, L.; and Niu, D. 2023. ConKI: Contrastive Knowledge Injection for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13610–13624.

Zhang, X.; Wei, W.; and Zou, S. 2025. Modal Feature Optimization Network with Prompt for Multimodal Sentiment Analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, 4611–4621.

Zhu, A.; Hu, M.; Wang, X.; Yang, J.; Tang, Y.; and Ren, F. 2024. KEBR: Knowledge Enhanced Self-Supervised Balanced Representation for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5732–5741.

Zhuang, Y.; Zhang, Y.; Hu, Z.; Zhang, X.; Deng, J.; and Ren, F. 2024. GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1800–1809.