SemMSA: Latent Semantic-Aided Robust Multimodal Sentiment Analysis with Incomplete Data

Wenhao Li[1,2], Zhibin Wu[1], Chong Xiao[1], Qiangchang Wang[1]*
[1]School of Software, Shandong University [2]Shenzhen Loop Area Institute

Abstract

Recent research on Multimodal Sentiment Analysis (MSA) has focused on learning from language, visual, and acoustic modalities with incomplete data to infer human sentiment. Most studies typically compensate for missing information by reconstructing modality features or designing complicated fusion mechanisms. However, these methods still suffer from spurious generation and noisy guidance due to the lack of high-level semantic grounding in partially observed multimodal evidence. To address these issues, we propose SemMSA, a latent semantic-aided framework that constructs rich sentiment-relevant semantics with LLMs, fully integrating with all modalities via anchor-free spectral alignment. It mainly consists of Cross-modal Semantic Refinement (CSR) and Cross-modal Spectral Alignment (CSA). Specifically, CSR first adaptively extracts visual and acoustic representations by corresponding adapters to form a unified multimodal prefix with language in the frozen LLM embedding space. It then iteratively produces continuous discriminative semantic states through a token-efficient latent refinement process without decoding explicit text. Next, CSA simultaneously aligns the refined semantics with all modalities by enhancing the dominant spectral component of their kernel Gram matrix. This captures global nonlinear dependencies among all representations without relying on a predefined anchor modality. In addition, an instance-level spectral separation constraint preserves cross-sample discriminability and mitigates representation collapse. Extensive experiments on SIMS, MOSI, and MOSEI benchmarks demonstrate that SemMSA achieves state-of-the-art performance. Project page: `https://peacelwh.github.io/SemMSA/`.

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to learn a comprehensive understanding of human sentiment by modeling multimodal information from language, vision, and audio [1, 2]. Recent research has increasingly moved from controlled laboratory settings toward real-world multimodal scenarios and deployment [3–6]. However, multimodal observations are often incomplete due to noise, occlusion, sensor failure, or transmission instability [7–10]. In particular, intra-modal missingness, where partial tokens, frames, or acoustic segments are corrupted within each modality, frequently occurs in practice and significantly degrades models trained under complete data assumptions.

Recent studies have made significant progress, following two paradigms, as shown in Figure 1(a) and (b). Reconstruction-based methods [11–16] aim to recover complete features from partially observed inputs, alleviating information loss. However, they tend to reconstruct low-level feature patterns rather than high-level semantics that are critical for sentiment understanding. Moreover, the same text may correspond to multiple plausible vocal prosodies and facial expressions, resulting in hallucinating features that contradict the actual sentiment evidence. Fusion-based methods [8, 9, 7, 17–19] instead directly integrate available modalities through complex cross-modal interaction modules to learn

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

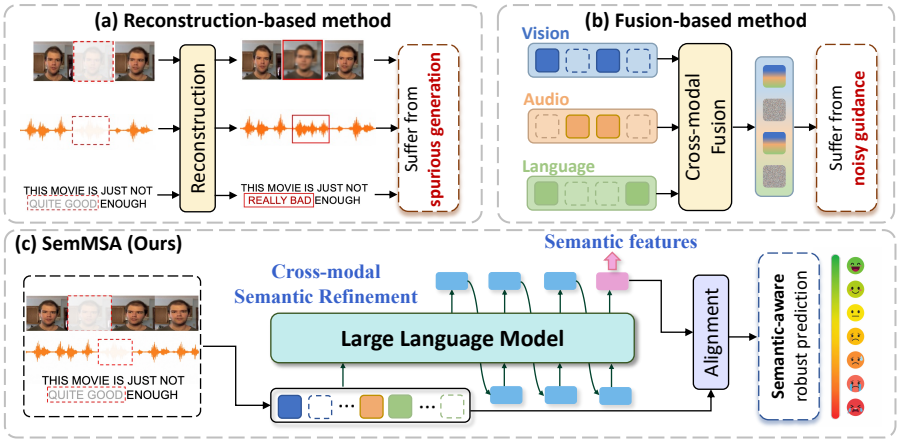

Figure 1: Framework comparison between (a) reconstruction-based methods, (b) fusion-based methods, and (c) our SemMSA. Compared with existing methods, SemMSA refines and fully integrates sentiment-relevant semantic information with the LLM to compensate for incomplete data.

robust representations. However, the fused features still suffer from noisy observations and limited evidence. They may overfit modality co-occurrence patterns observed during training under severe missing-modality scenarios, leading to unstable guidance and degraded discrimination.

Human sentiment understanding often relies on reliable contextual cues and prior knowledge about emotional expression to interpret intentions, attitudes, and affective states under incomplete observations [20, 21]. This motivates semantic-level compensation, where the goal is to enrich incomplete multimodal representations with rich sentiment-relevant semantic information. Recent studies also indicate a broader shift from discriminative emotion recognition toward generative emotion understanding [22]. Large Language Models (LLMs) [23–30], with strong contextual representation and semantic abstraction capabilities, provide a promising source of such high-level semantics. However, directly decoding textual descriptions or rationales introduces additional costly overhead. Therefore, an efficient mechanism is needed to exploit LLM-derived semantics in the hidden space.

Inspired by this perspective, we propose SemMSA, a latent semantic-aided framework with LLMs, as shown in Figure 1(c). First, Cross-modal Semantic Refinement (CSR) bridges heterogeneous non-language modalities and the frozen LLM embedding space through lightweight adapters. Each adapter uses learnable prompts with self- and cross-attention to adaptively aggregate reliable evidence from incomplete visual or acoustic sequences. The produced compact visual and acoustic prefix tokens are concatenated with language embedded in the frozen LLM space. CSR then recurrently appends the last hidden state of the LLM as a continuous latent token, yielding a sequence of sentiment-relevant semantic states. This provides token-efficient semantic compensation without explicit text decoding. Next, Cross-modal Spectral Alignment (CSA) establishes a kernel Gram matrix over the normalized semantic, language, visual, and acoustic representations for comprehensive cross-modal consistency alignment. It encourages all features to align simultaneously along a shared leading direction by enhancing the dominant spectral component of this matrix. This captures global nonlinear dependencies across representations, avoiding the instability from anchor modality under severe missingness. In addition, an instance-wise spectral separation constraint is imposed on the dominant eigenvectors to preserve cross-sample discriminability, mitigating representation collapse.

Overall, our contributions are summarized as follows:

- A SemMSA framework is proposed to provide high-level semantic compensation with LLMs and jointly align heterogeneous representations via an anchor-free spectral objective.
- CSR module adaptively aggregates multimodal data into the frozen LLM space and iteratively refines sentiment-relevant semantic states in a token-efficient latent process.
- CSA module aligns all representations simultaneously by enhancing the dominant spectral component they form, capturing global nonlinear relationships without anchor dependency.
- The proposed method consistently achieves state-of-the-art performance on MOSI, MOSEI, and SIMS under diverse missingness, significantly improving accuracy by 1.4% on average.

2 Related Work

2.1 MSA with Incomplete Data

Most MSA methods often assume that language, visual, and acoustic modalities are fully observed [31, 3, 32–34, 4, 35]. They learn unified representations by modeling intra- and inter-modal contextual dependencies. Once modalities are missing or corrupted, models trained under complete settings often degrade sharply. Recent fusion-based methods attempt to alleviate missing information through complex interaction and fusion, including multi-view correlation learning [36, 37], recurrent cross-modal translation [38, 39], graph-based dependency modeling [40], and modality-conditioned fusion [41]. Several dominance-guided approaches, such as LNLN [8] and P-RMF [9], further rely on high-quality primary modalities to anchor multimodal learning. Nevertheless, these methods may become unreliable under severe missingness [14]. The scarce representative features with noise in limited data may fail to capture sufficient discriminative information. Another line of work reconstructs missing modalities from observed ones, including cascaded residual auto-encoders [42], cycle-consistency-based recovery [38, 39], graph-based reconstruction [40], low-level feature restoration [12], and diffusion-based distribution recovery [14]. However, reconstruction-based methods mainly recover superficial statistical patterns rather than sentiment cues essential for sentiment judgment, while diffusion models also incur substantial computational costs. In contrast, SemMSA complements incomplete data with compact LLM-derived latent semantics and comprehensively integrates all representations through a consistent spectral alignment. To the best of our knowledge, this is the first attempt to utilize LLMs to perform semantic-level compensation for MSA.

2.2 Multimodal Alignment

Most existing methods [43–49] adopt traditional alignment or CLIP-based pairwise contrastive learning [50] . This paradigm has inspired a series of works extending to additional modalities, such as audio-to-text [51], point cloud-to-text [52], and video-to-text [53]. However, aligning each representation only to a single anchor neglects the interactions among the remaining points, making it difficult to capture global structural relationships. This inherently ignores the complexity and richness of multiple modalities. The Gram matrix, which characterizes the mutual geometry among sets of vectors, has shown promise in theoretical analyses of deep learning networks [54, 55] and various downstream tasks [56–61]. In contrast, SemMSA introduces a kernelized Gram matrix to capture nonlinear relationships, aligning all representations without anchor dependency.

3 Method

We begin by introducing the preprocessing of MSA, and then present two proposed components: (1) how cross-modal latent semantics are generated through CSR, and (2) how all representations are simultaneously aligned via CSA. Figure 2 shows the overview of the proposed SemMSA.

3.1 Multimodal Input

Given a multimodal sample for MSA, we consider three modalities of vision, audio, and language. Each modality with incomplete data is denoted by $\mathbf{U}_m$, where $m \in \{v, a, l\}$. Following prior works [8, 9, 11], each modality is first processed by a widely used modality-specific encoder f_ϕ with frozen parameters [62–64] to obtain a sequence representation $\mathbf{X}_m$ like other domains [65–68]:

$$\mathbf{X}_m = f_\phi(\mathbf{U}_m), \quad \mathbf{X}_m \in \mathbb{R}^{T_m \times d_m}, \tag{1}$$

where T_m is the sequence length and d_m denotes the feature dimension of each embedding vector.

3.2 Cross-modal Semantic Refinement

Visual and Audio Adapter. To extract heterogeneous nonverbal features into the embedding space of a frozen LLM, visual and acoustic adapters with M learnable prompt embeddings $\mathbf{P}_m \in \mathbb{R}^{M \times d_m}$ are introduced and updated by a lightweight transformer with B blocks. Due to the features $\mathbf{X}_m$ of frozen encoders without temporal information, learnable positional embeddings $\mathbf{t}_m \in \mathbb{R}^{T_m \times d_m}$ are injected into the frame or audio segment representations to obtain temporal order. $\mathbf{P}_m$ first act as

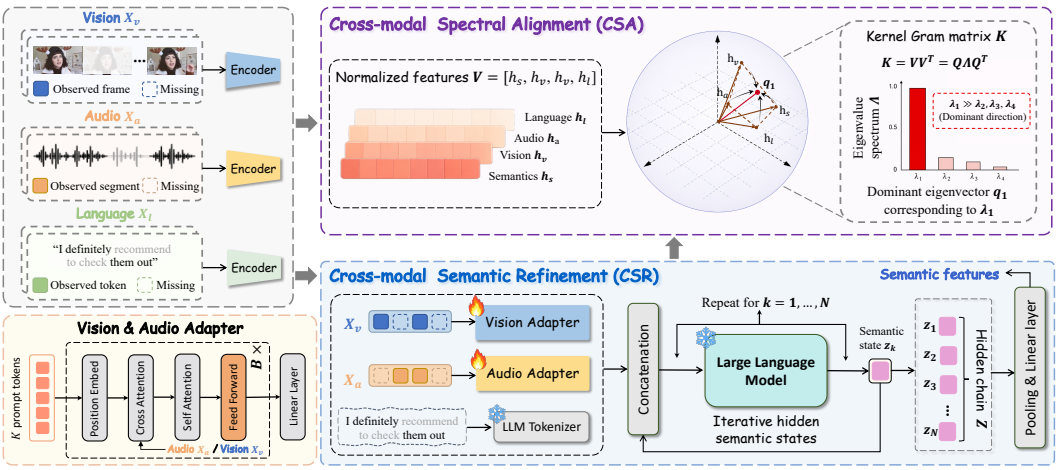

Figure 2: Overview of the proposed SemMSA framework. Multimodal inputs with incomplete data are first extracted by encoders. Then, CSR maps visual and audio features via adapters and concatenates with language embedded in the frozen LLM space. Iterative hidden-state refinement is performed to produce compact sentiment-aware latent semantic states. Next, CSA jointly aligns the semantics with visual, acoustic, and language representations by constructing a kernel Gram matrix over normalized features and enhancing its dominant spectral component.

queries to perform cross-attention with $\mathbf{X}_m$ to extract modality evidence. These embeddings then interact with each other to capture intra-modality dependencies through self-attention.

After B blocks, the output of the adapter consists of M modality-specific vectors, one per prompt embedding. They are linearly projected into the dimension of the LLM to obtain $\mathbf{Z}_v, \mathbf{Z}_a \in \mathbb{R}^{M\times d}$. The adapters are optimized to effectively bridge the output of the frozen encoders to the frozen LLM during training. In addition, the language $\mathbf{Z}_l \in \mathbb{R}^{T\times d}$ is directly embedded by the LLM, where T is the length of the textual embedding sequence and d is the LLM embedding dimension.

Semantic Refinement. The multimodal input $\mathbf{U}^{(0)} = [\mathbf{Z}_l; \mathbf{Z}_v; \mathbf{Z}_a] \in \mathbb{R}^{(T+2M)\times d}$ to the frozen LLM is formed by concatenating the three modalities in a fixed order. $\mathbf{U}^{(0)}$ serves as the initial multimodal prefix, providing the frozen LLM with partially observed language, visual, and acoustic evidence in a unified embedding space. Instead of decoding explicit textual descriptions, CSR performs iterative hidden-state refinement directly in the continuous LLM space. Let $\mathcal{F}_\theta(\mathbf{U}) \in \mathbb{R}^{|\mathbf{U}|\times d}$ denote the last-layer hidden states produced by the frozen LLM with parameters θ for an input sequence $\mathbf{U}$. The hidden state at the final position is defined as:

$$\mathbf{f}(\mathbf{U}) = \mathcal{F}_\theta(\mathbf{U})_{|\mathbf{U}|} \in \mathbb{R}^d, \tag{2}$$

where $\mathbf{f}(\mathbf{U})$ provides a contextual summary of the current multimodal prefix. Based on this representation, the k-th latent semantic state is constructed as

$$\mathbf{z}_k = \mathbf{f}\left(\mathbf{U}^{(k-1)}\right), \quad \mathbf{U}^{(k)} = \left[\mathbf{U}^{(k-1)}; \mathbf{z}_k\right], \quad k = 1, \ldots, O, \tag{3}$$

where O is the number of refinement steps, $\mathbf{z}_k \in \mathbb{R}^d$ denotes the continuous latent semantic state at step k, and $\mathbf{U}^{(k)} \in \mathbb{R}^{(T+2M+k)\times d}$ is the updated prefix after appending $\mathbf{z}_k$. Since $\mathbf{z}_k$ has the same dimensionality as the LLM token embeddings, it can be appended as a continuous input token for the next refinement step. We use this recurrent operation as a hidden-space semantic refinement mechanism, rather than explicit textual generation. The resulting sequence of latent semantic states is

$$\mathbf{Z} = [\mathbf{z}_1; \mathbf{z}_2; \cdots; \mathbf{z}_O] \in \mathbb{R}^{O\times d}. \tag{4}$$

These states provide a compact representation of sentiment-relevant information conditioned on the available multimodal evidence. By repeatedly applying the frozen LLM transformation to the updated prefix, CSR refines the hidden context while avoiding autoregressive decoding of natural-language descriptions. Finally, a compact sentiment-aware semantic representation $\mathbf{H}_\mathrm{s}$ is obtained for final fusion and prediction by applying pooling followed by a linear or two-layer MLP projection g_s:

$$\mathbf{H}_\mathrm{s} = g_\mathrm{s}\big(\mathrm{Pool}(\mathbf{Z})\big), \qquad \mathbf{H}_\mathrm{s} \in \mathbb{R}^{d'}. \tag{5}$$

3.3 Cross-modal Spectral Alignment

To achieve comprehensive consistency alignment among semantics, vision, audio, and language, Cross-modal Spectral Alignment (CSA) is proposed to improve global and nonlinear interactions among all representations. Unlike traditional pairwise contrastive objectives which align modality pairs through a predefined anchor, CSA measures the alignment by analyzing the spectral structure of their shared representation matrix. A stronger dominant spectral component indicates that different modalities are concentrated along a common latent direction.

Specifically, three modality representations $\mathbf{X}_m$ are first extracted and unified by their respective modality encoders to obtain $\mathbf{H}_m$ with the d' dimension. Each encoder consists of a linear transformation layer followed by Transformer encoder layers. Next, they are transformed into normalized quadruples and construct the matrix $\mathbf{V} = (\mathbf{h}_s, \mathbf{h}_v, \mathbf{h}_a, \mathbf{h}_l)$. The corresponding Gram matrix $\mathbf{G} \in \mathbb{R}^{4\times4}$ is first defined to reflect the pairwise similarity as follows:

$$\mathbf{G}(\mathbf{V}) = \mathbf{V}^\top\mathbf{V} = \begin{bmatrix} \langle\mathbf{h}_s,\mathbf{h}_s\rangle & \langle\mathbf{h}_s,\mathbf{h}_v\rangle & \langle\mathbf{h}_s,\mathbf{h}_a\rangle & \langle\mathbf{h}_s,\mathbf{h}_l\rangle \\ \langle\mathbf{h}_v,\mathbf{h}_s\rangle & \langle\mathbf{h}_v,\mathbf{h}_v\rangle & \langle\mathbf{h}_v,\mathbf{h}_a\rangle & \langle\mathbf{h}_v,\mathbf{h}_l\rangle \\ \langle\mathbf{h}_a,\mathbf{h}_s\rangle & \langle\mathbf{h}_a,\mathbf{h}_v\rangle & \langle\mathbf{h}_a,\mathbf{h}_a\rangle & \langle\mathbf{h}_a,\mathbf{h}_l\rangle \\ \langle\mathbf{h}_l,\mathbf{h}_s\rangle & \langle\mathbf{h}_l,\mathbf{h}_v\rangle & \langle\mathbf{h}_l,\mathbf{h}_a\rangle & \langle\mathbf{h}_l,\mathbf{h}_l\rangle \end{bmatrix}, \quad \mathbf{G}_{ij} = \langle\mathbf{h}_i,\mathbf{h}_j\rangle. \tag{6}$$

To further learn nonlinear cross-modal dependency, this metric is extended into a high-dimensional Reproducing Kernel Hilbert Space (RKHS) using a Radial Basis Function (RBF) kernel mapping $\kappa(\cdot,\cdot)$ The corresponding kernel Gram matrix $\mathbf{K}(V)$ is defined as:

$$\mathbf{K}_{ij} = \kappa(\mathbf{h}_i,\mathbf{h}_j), \qquad \kappa(\mathbf{h}_i,\mathbf{h}_j) = \exp\left(-\frac{\|\mathbf{h}_i-\mathbf{h}_j\|_2^2}{2\sigma^2}\right). \tag{7}$$

If the four representations of the same instance are sufficiently aligned in the kernel space, then the kernel Gram matrix should exhibit a significant low-rank structure. In the ideal case, all representations are mapped to the same shared semantic direction, and $\mathbf{K}$ degenerates into a rank-one matrix. Therefore, we perform eigendecomposition on $\mathbf{K}$ as follows:

$$\mathbf{K} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top, \qquad \mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\lambda_2,\lambda_3,\lambda_4), \tag{8}$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \geq 0$. The largest eigenvalue λ_1 represents the principal nonlinear semantic direction shared by the four representations. A higher proportion of it indicates that the representations are more concentrated in the same latent alignment subspace. Therefore, we regard the eigenvalues as logits and enhance the dominance of λ_1 through a softmax-based spectral objective:

$$\mathcal{L}_{\mathrm{csa}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\lambda_1^i/\tau)}{\sum_{j=1}^{4}\exp(\lambda_j^i/\tau)}, \tag{9}$$

where N is the batch size and τ denotes the temperature parameter. To prevent degenerate solutions where different instances collapse to the same semantic point, we introduce an instance-wise spectral separation regularizer. For the i-th instance, the dominant eigenvector $\mathbf{q}_1^i$ associated with the largest eigenvalue λ_1^i of the kernel Gram matrix indicates the principal alignment pattern among the semantic, visual, acoustic, and language representations. Since $\mathbf{q}_1^i$ only reflects representation-wise combination coefficients, we project it back to the high-dimensional representation space to obtain the dominant semantic direction $\mathbf{u}_1^i$. The separation loss is then defined as:

$$\mathbf{u}_1^i = \frac{\mathbf{V}^i\mathbf{q}_1^i}{\left\|\mathbf{V}^i\mathbf{q}_1^i\right\|_2}, \qquad \mathcal{L}_{\mathrm{sep}} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\left[(\mathbf{u}_1^i)^\top\mathbf{u}_1^j\right]^2. \tag{10}$$

where $\mathbf{V}^i$ contains the normalized representations of the i-th instance. This regularizer penalizes high similarity between dominant semantic directions of different instances, preserving inter-instance discriminability. These aligned embeddings are fused via element-wise summation to form the final multimodal representation $\hat{\mathbf{H}}$ as follows:

$$\hat{\mathbf{H}} = \mathbf{H}_s + \mathbf{H}_v + \mathbf{H}_a + \mathbf{H}_l. \tag{11}$$

Table 1: The average results on MOSI and MOSEI datasets over missing rates from 0.0 to 0.9.

	MOSI						MOSEI					
Method	Acc-2	F1	Acc-5	Acc-7	MAE ↓	Corr ↑	Acc-2	F1	Acc-5	Acc-7	MAE ↓	Corr ↑
MISA [69]	70.33/71.49	70.00/71.28	33.08	29.85	1.085	0.524	75.82/71.27	68.73/63.85	39.39	40.84	0.780	0.503
Self-MM [70]	69.26/70.51	67.54/66.60	34.67	29.55	1.070	0.512	77.42/73.89	72.31/68.92	45.38	44.70	0.695	0.498
MMIM [71]	67.06/69.14	64.04/66.65	33.77	31.30	1.077	0.507	75.89/73.32	70.32/68.72	41.74	40.75	0.739	0.489
TETFN [72]	67.68/69.76	63.29/65.69	34.34	30.30	1.087	0.507	-	-	-	-	-	-
TFR-Net [11]	66.35/68.15	60.06/61.73	34.67	29.54	1.200	0.459	77.23/73.62	71.99/68.80	34.67	46.83	0.697	0.489
ALMT [4]	68.39/70.40	71.80/72.57	33.42	30.30	1.083	0.498	77.54/76.64	78.03/77.14	41.64	40.92	0.674	0.481
LNLN [8]	70.94/72.55	71.25/72.73	38.27	34.26	1.046	0.527	78.19/76.30	79.95/77.77	46.17	45.42	0.692	0.530
P-RMF [9]	71.53/72.81	71.69/72.93	38.50	34.19	1.038	0.525	78.83/78.14	80.39/79.33	45.87	44.63	0.658	0.589
TF-Mamba [73]	73.46/72.54	73.59/72.57	37.74	33.95	1.035	0.548	77.34/77.61	77.18/77.43	46.64	45.66	0.673	0.578
SemMSA	**74.36/73.91**	**74.18/73.82**	**40.93**	**36.49**	**1.011**	**0.550**	**79.61/79.38**	**80.62/79.87**	**48.12**	**47.06**	**0.648**	**0.601**

The fused feature $\hat{\mathbf{H}}$ is finally fed into a linear or MLP regressor $\mathcal{C}$ for sentiment prediction $\hat{y} = \mathcal{C}(\hat{\mathbf{H}})$. The overall loss uses a regression loss ℓ (MSE or L1) on $\hat{y}$ and on the semantic prediction $\mathcal{C}(\mathbf{H}_s)$ with weight λ_s:

$$\mathcal{L}_{\text{total}} = \frac{1}{N}\sum_{i=1}^{N}\Big[\ell\big(y^i, \hat{y}^i\big) + \lambda_s\,\ell\big(y^i, \mathcal{C}(\mathbf{H}_s^i)\big)\Big] + \mathcal{L}_{\text{csa}} + \mathcal{L}_{\text{sep}}. \tag{12}$$

4 Experiments

4.1 Experimental Details

Datasets. Extensive experiments are conducted on three standard MSA benchmarks: MOSI [74] and MOSEI [75] annotated with sentiment scores in $[-3, +3]$. SIMS [76] is a Chinese dataset annotated in $[-1, +1]$. The detailed dataset statistics are introduced in the Appendix.

Implementation Details. SemMSA is trained with a 200-epoch schedule using AdamW [77], warm-up, and cosine annealing. The batch size (64 or 32) and learning rate (5×10^{-5} to 2×10^{-4}) are set per metric, as listed in the Appendix. The input sequence length T is 8, and the hidden dimension d is 128. Qwen3-1.7B [78] is adopted as the frozen LLM to generate semantic features. The visual and acoustic Adapters use $M = 8$ learnable prompts and $B = 2$ blocks. Refinement step O in CSR is set to 4. The RBF kernel bandwidth $\sigma = 1.0$ and temperature $\tau = 1.0$ by default. All experiments are performed with NVIDIA RTX 6000 Ada and RTX 4090 GPUs.

Missingness Settings and Evaluation Metrics. Following [8, 9, 11, 73], missing visual and acoustic segments are replaced with zero vectors and language tokens with `[UNK]`. Training employs instance-wise Bernoulli masking (see Table 9). During testing, the missing rate is varied from 0 to 0.9 (step 0.1) and missing positions are independently sampled per modality. Random seeds 1111, 1112, and 1113 are used, as detailed in the Appendix. For MOSI and MOSEI, we report seven-class accuracy (Acc-7), Acc-5, Acc-2, F1 score, Mean Absolute Error (MAE), and Pearson correlation coefficient (Corr). Acc-2 and F1 are reported sequentially in two forms of negative/non-negative and negative/positive. For SIMS, we report Acc, F1 score, MAE, and Corr.

4.2 Main Results

Intra-modal Missingness. Table 1 and 2 report comparisons on MOSI, MOSEI, and SIMS under intra-modal missing settings. SemMSA consistently achieves state-of-the-art performance. On MOSI, it improves Acc-5 and Acc-7 over the best baselines TF-Mamba [73] by 2.43% and 2.23%, respectively, while achieving the lowest MAE of 1.011. On MOSEI, SemMSA further improves Acc-2 to 79.61 and F1 to 80.62, demonstrating its scalability on larger-scale data. On SIMS, the improvement is

Table 2: The average results on SIMS dataset over missing rates from 0.0 to 0.9.

Method	Acc-2	F1	Acc-3	Acc-5	MAE ↓	Corr ↑
MISA [69]	72.71	66.30	56.87	31.53	0.539	0.348
Self-MM [70]	72.81	68.43	56.75	32.28	0.508	0.376
MMIM [71]	69.86	66.21	56.76	31.81	0.544	0.339
TETFN [72]	73.58	68.67	56.91	33.42	0.505	0.387
TFR-Net [11]	68.13	58.70	52.89	26.52	0.661	0.169
ALMT [4]	71.85	76.21	56.47	34.16	0.509	0.372
LNLN [8]	73.15	69.18	57.32	34.68	0.515	0.387
P-RMF [9]	73.64	74.65	54.75	34.83	0.500	0.414
TF-Mamba [73]	74.68	72.20	55.51	34.46	0.512	0.386
SemMSA	**75.46**	**77.50**	**58.84**	**35.68**	**0.474**	**0.501**

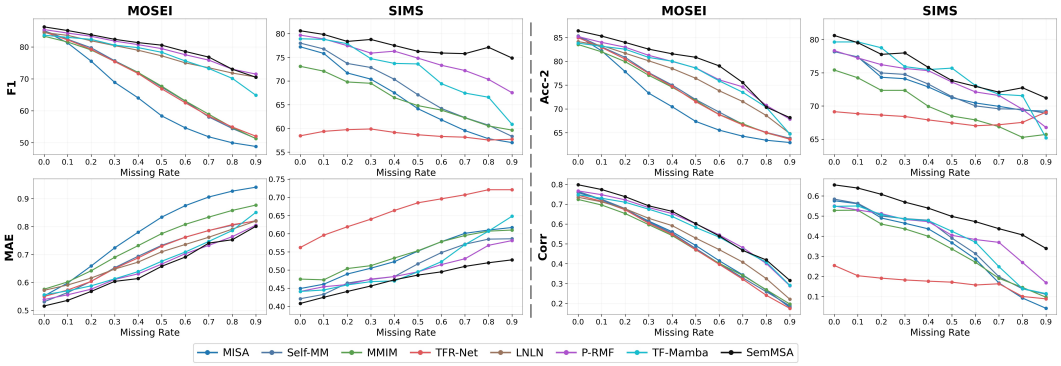

Figure 3: Performance visualization of F1 Score, MAE, Acc-2, and Corr across MOSI, MOSEI, and SIMS under missing rates from 0.0 to 0.9, where lower MAE denotes superior performance.

particularly clear, surpassing P-RMF [9] by 0.087 in Corr. Figure 3 further illustrates the performance trends of F1, Acc-2, MAE, and Corr. As the missing rate increases, most baseline methods exhibit clear performance degradation, struggling to maintain reliable representations. In contrast, SemMSA maintains consistently superior and more stable performance across missing rates. These results indicate that SemMSA can effectively complement incomplete multimodal observations with sentiment-relevant latent semantics and comprehensively align heterogeneous representations through global nonlinear alignment, leading to strong robustness and generalization under challenging missing-modality scenarios. Detailed results under each missing rate are provided in the Appendix.

Inter-modal Missingness. Following the same protocol in [17, 40, 80], we further remove entire modalities from MOSEI samples to evaluate the performance of SemMSA under inter-modal missingness. The results are reported in Table 3 using F1 as the evaluation metric. It is worth noting that (1) all methods degrade significantly in unimodal scenarios, confirming the complementary roles of different modalities. (2) Language consistently outperforms visual and acoustic modalities, highlighting more explicit sentiment semantics in textual information. (3) SemMSA achieves the best results under all inter-modal missingness, outperforming the second-best CorrKD [17] by a significant margin of 2.71% on average. Improvements in the language missingness, such as {a,v}, demonstrate that SemMSA effectively leverages nonverbal evidence to obtain rich sentiment-relevant semantics by semantic refinement, fully integrating them via anchor-free spectral alignment.

Table 3: The F1 scores on the MOSEI dataset under varying inter-modal missing conditions.

Method	$\{l\}$	$\{a\}$	$\{v\}$	$\{l,a\}$	$\{l,v\}$	$\{a,v\}$	$\{l,a,v\}$	Avg.
Self-MM [70]	71.53	43.57	37.61	75.91	74.62	49.52	83.69	62.35
CubeMLP [79]	67.52	39.54	32.58	71.69	70.06	48.54	83.17	59.01
DMD [80]	70.26	46.18	39.84	74.78	72.45	52.70	84.78	63.00
MCTN [38]	75.50	62.72	59.46	76.64	77.13	64.84	81.75	71.15
TransM [81]	77.98	63.68	58.67	80.46	78.61	62.24	81.48	71.87
SMIL [82]	76.57	65.96	60.57	77.68	76.24	66.87	80.74	72.09
GCNet [40]	80.52	66.54	61.83	81.96	81.15	69.21	82.35	74.79
CorrKD [17]	80.76	66.09	62.30	81.74	81.28	71.92	82.16	75.18
SemMSA	**83.61**	**66.59**	**65.91**	**85.17**	**84.96**	**72.68**	**86.32**	**77.89**

Table 4: Ablation study of components and losses on MOSI and SIMS datasets.

CSR	$\mathcal{L}_{csa}$	$\mathcal{L}_{sep}$	MOSI						SIMS					
			Acc-2	F1	Acc-5	Acc-7	MAE ↓	Corr ↑	Acc-2	F1	Acc-3	Acc-5	MAE ↓	Corr ↑
			67.49/69.47	63.28/67.01	34.29	31.71	1.085	0.501	66.83	57.92	31.12	54.61	0.593	0.387
✓			73.36/73.29	73.04/73.28	38.84	35.03	1.040	0.522	74.18	76.21	34.27	57.36	0.488	0.450
✓	✓		73.94/73.56	73.83/73.46	40.48	36.08	1.024	0.529	74.92	76.88	35.02	58.12	0.480	0.493
✓	✓	✓	**74.36/73.91**	**74.18/73.82**	**40.93**	**36.49**	**1.011**	**0.550**	**75.46**	**77.50**	**35.68**	**58.84**	**0.474**	**0.501**

4.3 Model Analysis

Ablation Study. We conduct an ablation study on MOSI and SIMS to evaluate the effectiveness of the proposed CSR, $\mathcal{L}_{csa}$, and $\mathcal{L}_{sep}$, as shown in Table 4. First, introducing CSR alone brings substantial gains over the baseline on both datasets, improving Acc-2 and F1 by 5.68% and 11.11% on average. These indicate that rich sentiment-relevant semantics effectively compensate for incomplete

Table 5: Comparison with different projection and alignment methods on SIMS dataset.

(a) Projection mechanisms

Method	Acc-2	F1	Acc-5	Acc-3	MAE ↓	Corr ↑
Linear	72.84	73.62	54.91	32.46	0.512	0.431
MLP	73.58	74.36	56.03	33.42	0.501	0.452
Trans. [83]	74.52	76.21	57.63	34.76	0.486	0.482
QFormer [84]	74.83	76.68	58.12	35.09	0.481	0.491
Ours	**75.46**	**77.50**	**58.84**	**35.68**	**0.474**	**0.501**

(b) Alignment mechanisms

Method	Acc-2	F1	Acc-5	Acc-3	MAE ↓	Corr ↑
InfoNCE [85]	73.42	75.39	56.88	33.74	0.501	0.489
CMD [69]	73.77	75.18	56.92	34.21	0.503	0.487
PMRL [58]	74.63	75.98	57.62	34.58	0.493	0.499
Volume [57]	74.31	76.35	57.91	34.86	0.487	0.506
Ours	**75.46**	**77.50**	**58.84**	**35.68**	**0.474**	**0.501**

inputs. Second, adding $\mathcal{L}_{\text{csa}}$ further brings performance gains. MOSI Acc-5 increases from 38.84 to 40.48 and Acc-7 from 35.03 to 36.08, suggesting better cross-modal consistency. Finally, the best performance is achieved when $\mathcal{L}_{sep}$ is also incorporated. The full model consistently outperforms all ablated variants, highlighting the importance of semantic refinement for enhancing discriminative sentiment semantics, as well as the complementarity between intra-instance cross-modal coherence by $\mathcal{L}\text{csa}$ and inter-instance discriminability by $\mathcal{L}\text{sep}$.

Comparison with Projection Methods. As shown in Table 5a, we compare different projection strategies by replacing our Adapter with representative alternatives. Directly projecting modality features into the LLM embedding space with linear or MLP leads to a clear performance drop, indicating that dimensional matching alone is insufficient for constructing effective multimodal prefixes. Despite the effectiveness of Standard Transformer ($\sim$85M) [83] and Q-Former ($\sim$100M) [84], they introduce substantially higher computational overhead. In contrast, our lightweight Adapter with only 4.7M parameters achieves the best overall results, indicating that it can compactly aggregate sequence-level modal evidence and adaptively select reliable clues under incomplete observations.

Comparison with Alignment Methods. As shown in Table 5b, we compare four cross-modal alignment strategies. The baseline is utilized without an alignment objective. InfoNCE [85] and CMD [69] rely on pairwise similarity, failing to model the joint structure across all three modalities, resulting in poor results. While the PMRL [58] and volume-based loss improve upon this, they remain limited to the linear space and anchor modality, struggling to capture complex semantic relationships and select a stable anchor under severe intra-modality missingness. In contrast, our method introduces a Reproducing Kernel Hilbert Space (RKHS) without anchor dependency to enable nonlinear and robust alignment, achieving the best results.

Table 6: Efficiency comparison on MOSI.

Method	Trainable Params	Task GFLOPs	Memory(GB)	Latency(ms)	Acc-2(%)	F1(%)
LNLN [8]	116M	9.0	12.8	25.1	70.94/72.55	71.25/72.73
P-RMF [9]	117M	9.7	13.5	67.0	71.53/72.81	71.69/72.93
SemMSA	113M	4.8	13.0	30.6	74.36/73.91	74.18/73.82

Comparison of Computational Overhead. Efficiency experiments are further conducted on MOSI using the same server configuration as LNLN [8] and P-RMF [9], where trainable parameters and GFLOPs include only the modules optimized for MSA. As shown in Table 6, SemMSA achieves the best performance while maintaining a comparable computation cost. It reduces the task GFLOPs by 50.5% and inference time by 54.3% compared with P-RMF. Notably, the reported memory and latency reflect the entire pipeline with the frozen LLM, rather than just the trainable modules. The efficiency mainly comes from the token-efficient hidden-state refinement design. SemMSA avoids autoregressive decoding of explicit textual descriptions and appends only a small number of continuous latent states. These results indicate that SemMSA provides a favorable trade-off between robustness and inference cost under incomplete multimodal settings.

Visualization of Prediction Performance. To further evaluate the robustness of SemMSA, we visualize the confusion matrices on the MOSI dataset under missing rates of 0.1, 0.5, 0.7, and 0.9, as shown in Figure 4. Under high missing rates, LNLN tends to wrongly concentrate predictions

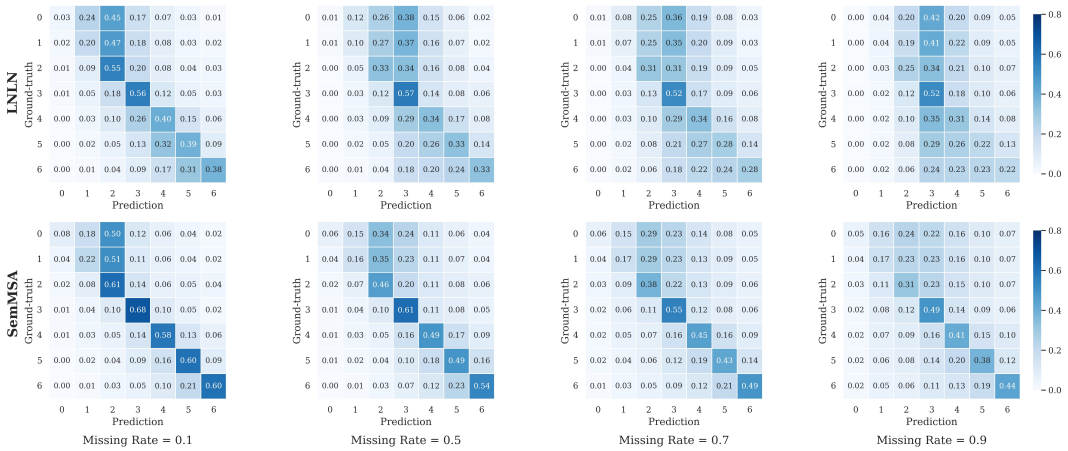

Figure 4: Confusion matrices of SemMSA and LNLN on the MOSI, where 0-6 denote strongly negative, negative, weakly negative, neutral, weakly positive, positive, and strongly positive, respectively.

around middle sentiment classes. In contrast, SemMSA maintains clearer diagonal patterns across sentiment categories. This demonstrates that the semantic-aware modeling in SemMSA facilitates the capture of discriminative sentiment cues from incomplete inputs, alleviating classification boundary degeneration for robust prediction under severe modality missingness.

Table 7: Comparison with LLMs for prediction and semantic generation on MOSI and SIMS.

Model	Role	MOSI						SIMS					
		Acc-2	F1	Acc-5	Acc-7	MAE ↓	Corr ↑	Acc-2	F1	Acc-5	Acc-3	MAE ↓	Corr ↑
Qwen2.5-Omni-7B [86]	Prediction	69.84/71.36	69.92/71.58	36.94	33.21	1.086	0.497	70.83	70.26	52.74	31.92	0.548	0.341
Qwen3-1.7B [78]	Semantics	74.36/73.91	74.18/**73.82**	40.93	36.49	1.011	0.550	75.46	77.50	58.84	35.68	0.474	0.501
Llama3.1-8B [87]	Semantics	74.42/73.75	74.04/73.69	**41.11**	36.50	1.015	0.546	-	-	-	-	-	-
Qwen3-8B [78]	Semantics	**74.66**/73.88	74.11/73.79	40.86	**36.62**	1.012	**0.553**	75.38	**77.63**	58.79	**35.77**	0.476	**0.514**
Qwen2.5-Omni-7B [86]	Semantics	74.52/**74.03**	**74.26**/73.76	41.05	36.55	**1.008**	0.548	**75.59**	77.42	**58.96**	35.61	**0.471**	0.512

The Effect of Different LLMs. Table 7 compares open-source Qwen3-1.7B, Llama-3.1-8B, Qwen3-8B, and Qwen2.5-Omni-7B on MOSI and SIMS. For fair comparison, the direct prediction of LLMs strictly follows the same evaluation protocol and data splits for MSA with incomplete data. However, its performance is relatively limited with only 69.84/71.36 Acc-2 and 69.92/71.58 F1 on MOSI, showing the difficulty of adapting LLMs to sentiment prediction. In contrast, using LLMs for semantic generation brings clear improvements, since rich semantic information provides auxiliary high-level knowledge to compensate for scarce representations with missing data. Specifically, Qwen3-8B achieves the best F1 and Acc-3 on SIMS. Llama3.1-8B without support of Chinese and Qwen2.5-Omni-7B obtains superior results of Acc-5 and MAE on MOSI, respectively. These results show that SemMSA is robust across different LLMs and that LLMs are more effective as semantic generators than direct predictors for MSA with incomplete data.

5 Conclusion

In this paper, we propose SemMSA, a semantic-aided framework to enrich incomplete data with LLM-derived sentiment-aware semantic representations and integrate all modalities through anchor-free spectral alignment. Specifically, we introduce Cross-modal Semantic Refinement (CSR), which maps visual and acoustic representations into the frozen LLM embedding space with lightweight Adapters. CSR then produces sentiment-relevant latent semantics iteratively through token-efficient hidden-state refinement, without decoding explicit text. We further propose Cross-modal Spectral Alignment (CSA), which jointly aligns semantic, language, visual, and acoustic representations by enhancing the dominant spectral component of their kernel Gram matrix. This captures global nonlinear cross-modal dependencies without relying on a predefined anchor modality, while preserving discriminative structure through instance-level spectral separation. Extensive experiments on SIMS, MOSI, and MOSEI datasets demonstrate the effectiveness of SemMSA, achieving state-of-the-art performance under diverse missing-modality settings.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China under Grant 62406177, in part by the Shandong Excellent Young Scientists Fund (Oversea) under Grant 2024HWYQ-027, in part by the Natural Science Foundation of Shandong province, China under Grant ZR2023QF124, in part by the Young Scholars Program of Shandong University.

References

[1] Y. Fang, W. Huang, G. Wan, K. Su, and M. Ye, "Emoe: Modality-specific enhanced dynamic emotion experts," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 314–14 324.

[2] S. Wu, D. He, X. Wang, L. Wang, and J. Dang, "Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 2, 2025, pp. 1601–1609.

[3] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 6558–6569.

[4] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[5] P. Wang, Q. Zhou, Y. Wu, T. Chen, and J. Hu, "Dlf: Disentangled-language-focused multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 21 180–21 188.

[6] H. Xu, T. Li, W. Chen, Y. Liu, X. Zuo, Y. Song, and H. Li, "Enhancing vision-language navigation with multimodal event knowledge from real-world indoor tour videos," *arXiv preprint arXiv:2602.23937*, 2026.

[7] M. Li, D. Yang, Y. Lei, S. Wang, S. Wang, L. Su, K. Yang, Y. Wang, M. Sun, and L. Zhang, "A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 9, 2024, pp. 10 074–10 082.

[8] H. Zhang, W. Wang, and T. Yu, "Towards robust multimodal sentiment analysis with incomplete data," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 943–55 974, 2024.

[9] A. Zhu, M. Hu, X. Wang, J. Yang, Y. Tang, and N. An, "Proxy-driven robust multimodal sentiment analysis with incomplete data," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 22 123–22 138.

[10] W. Li, Q. Wang, J. Li, S. Zhao, M. Ruan, and Y. Yin, "Semantic-aware adaptation with hierarchical multimodal prompts for few-shot learning," in *2025 IEEE International Conference on Multimedia and Expo*. IEEE, 2025, pp. 1–6.

[11] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4400–4407.

[12] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 309–325, 2023.

[13] J. Zeng, T. Liu, and J. Zhou, "Tag-assisted multimodal sentiment analysis under uncertain missing modalities," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1545–1554.

[14] Y. Wang, Y. Li, and Z. Cui, "Incomplete multimodality-diffused emotion recognition," *Advances in Neural Information Processing Systems*, vol. 36, pp. 17 117–17 128, 2023.

[15] B. Pu, X. Liang, X. Ding, J. Wu, Z. Lei, S. Li, K. Li, and J. Ma, "F^2-assist: Multi-phase fetal growth forecast and report generation from ultrasound examination," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 35 341–35 350.

[16] H. Xu, Y. Liu, T. Li, R. Shen, B. Jiang, J. Peng, D. Luo, X. Hu, S. Yan, and H. Li, "Irpo: Boosting image restoration via post-training grpo," *arXiv preprint arXiv:2512.00814*, 2025.

[17] M. Li, D. Yang, X. Zhao, S. Wang, Y. Wang, K. Yang, M. Sun, D. Kou, Z. Qian, and L. Zhang, "Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 458–12 468.

[18] D. Kim and T. Kim, "Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models," in *European Conference on Computer Vision*. Springer, 2024, pp. 171–187.

[19] Z. Wu, W. Li, Z. Qin, and Q. Wang, "Expanding intra-modal and compressing inter-modal embedding space for incomplete multimodal emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[20] M. C. Camacho, A. N. Nielsen, D. Balser, E. Furtado, D. C. Steinberger, L. Fruchtman, J. P. Culver, C. M. Sylvester, and D. M. Barch, "Large-scale encoding of emotion concepts becomes increasingly similar between individuals from childhood to adolescence," *Nature neuroscience*, vol. 26, no. 7, pp. 1256–1266, 2023.

[21] J. Ortega, Y. Murai, and D. Whitney, "Integration of affective cues in context-rich and dynamic scenes varies across individuals," *Nature Communications*, vol. 17, p. 786, 2025.

[22] Z. Lian, X. Peng, K. Xu, Z. Jia, X. Che, Z. Cheng, F. Ma, L. Cui, Y. Zhang, X. Liu *et al.*, "Mer 2026: From discriminative emotion recognition to generative emotion understanding," *arXiv preprint arXiv:2604.19417*, 2026.

[23] Y. Liu, Y. Liu, X. Yin, B. Wang, C. Zhang, H. Yin, and Z. Han, "Openclawbench: Benchmarking process-side anomalies in real-world agent execution trajectories," *arXiv preprint arXiv:2605.29253*, 2026.

[24] J. Wu, X. Liang, X. Bai, and Z. Chen, "Surgbox: Agent-driven operating room sandbox with surgery copilot," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 2041–2048.

[25] Q. Zhang, Z. Han, H. Sun, Y. Gong, X. Wang, Y. Yin, and S. Li, "Stabilizing feature geometry in noisy pretrained models for robust downstream tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 789–800.

[26] H. Xu, T. Li, W. Chen, Y. Liu, J. Wu, H. Lei, Y. Lou, L. Wang, H. Wang, and H. Li, "Hcsg: Human-centric semantic-geometric reasoning for vision-language navigation," *arXiv preprint arXiv:2605.13321*, 2026.

[27] M. Yang, J. Piao, X. Xia, X. Lan, J. Chen, Y. Gong, and Y. Li, "Skillmaster: Toward autonomous skill mastery in llm agents," *arXiv preprint arXiv:2605.08693*, 2026.

[28] X. Xia, J. Piao, M. Yang, X. Lan, J. Chen, and Y. Li, "From scoring to acting: Outcome-verified comparative self-distillation for llm agents," *arXiv preprint arXiv:2607.27937*, 2026.

[29] M. Yang, Y. Pan, J. Piao, D. Song, Y. Gong, and Y. Li, "Enterprisebench: Benchmarking llm agents on enterprise-level strategic reasoning and decision-making," *arXiv preprint arXiv:2609.37658*, 2026.

[30] W. Li, Q. Wang, P. Zhao, and Y. Yin, "Knn transformer with pyramid prompts for few-shot learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1082–1091.

[31] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 1103–1114.

[32] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Transactions on Multimedia*, vol. 22, pp. 122–137, 2019.

[33] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2554–2562.

[34] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1642–1651.

[35] H. Zhang, Y. Zhang, C. Ying, X. Tang, and T. Yu, "Improving task-specific multimodal sentiment analysis with general mllms via prompting," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[36] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[37] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 2015, pp. 1083–1092.

[38] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 6892–6899.

[39] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2608–2618.

[40] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "Gcnet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 45, no. 7, pp. 8419–8432, 2023.

[41] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, vol. 101, p. 101973, 2024.

[42] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1405–1414.

[43] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[44] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned gradient for prompt tuning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 15 659–15 669.

[45] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[46] L. Meng, Z. Qi, L. Wu, X. Du, Z. Li, L. Cui, and X. Meng, "Improving global generalization and local personalization for federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 76–87, 2024.

[47] Z. Qi, Y. Pan, L. Meng, S. Zhou, H. Yu, X. Li, and X. Meng, "Global prompt refinement with non-interfering attention masking for one-shot federated learning," *Advances in Neural Information Processing Systems*, vol. 38, pp. 12 121–12 145, 2026.

[48] Z. Qi, L. Meng, Z. Li, H. Hu, and X. Meng, "Cross-silo feature space alignment for federated learning on clients with imbalanced data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 19, 2025, pp. 19 986–19 994.

[49] W. Li, X. Meng, Q. Wang, Z. Han, Z. Wu, and Y. Yin, "Dvla-rl: Dual-level vision-language alignment with reinforcement learning gating for few-shot learning," in *ICLR*, 2026.

[50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[51] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[52] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8552–8562.

[53] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 638–647.

[54] J. Pennington and P. Worah, "Nonlinear random matrix theory for deep learning," *Advances in neural information processing systems*, vol. 30, 2017.

[55] W. Li, Q. Wang, X. Meng, Z. Wu, and Y. Yin, "Vt-fsl: Bridging vision and text with llms for few-shot learning," in *NeurIPS*, vol. 38, 2025, pp. 113 697–113 727.

[56] I. Nejjar, Q. Wang, and O. Fink, "Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 744–11 754.

[57] G. Cicchetti, E. Grassucci, L. Sigillo, and D. Comminiello, "Gramian multimodal representation learning and alignment," in *The Thirteenth International Conference on Learning Representations*, 2025.

[58] X. Liu, X. Xia, S.-K. Ng, and T.-S. Chua, "Principled multimodal representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[59] P. Wang, S. Sun, J. Lu, C. Wu, Z. Zhang, Y. Zhang, C. Shan, D. Lu, G. Zhang, and Z. Zheng, "Cast-lut: tokenizer-guided hsv look-up tables for purple flare removal," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 12, 2026, pp. 10 002–10 010.

[60] X. Liang, L. Zhou, N. Li, M. Xu, Z. Song, D. Yi, J. Wu, J. Ma, H. Liu, Z. Lei *et al.*, "Multimodal causality-driven representation learning for generalizable medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 13 670–13 679.

[61] Z. Wang, W. Zhang, W. Zhang, F. Zhang, X. Tan, Y. Qin, and G. Li, "Navigation world models as group actions," in *Advances in Neural Information Processing Systems*, 2026.

[62] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[63] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." *SciPy*, vol. 2015, pp. 18–24, 2015.

[64] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[65] Z. Wang, W. Zhang, W. Zhang, X. Tan, J. Feng, X. Luo, and G. Li, "Lanediffusion++: Feature-level generative priors for structured lane graph learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[66] Z. Wang, W. Zhang, W. Zhang, X. Tan, H. Liu, Y. Wang, and G. Li, "Lanediffusion: Improving centerline graph learning via prior injected bev feature generation," in *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025, pp. 27 052–27 062.

[67] R. He and J. Shi, "Prior-free tabular test-time adaptation," in *International Conference on Learning Representations*, vol. 2026, 2026, pp. 69 478–69 494.

[68] R. He, Y. Dong, L.-Z. Guo, Y. Yin, and T. Wu, "Re-evaluating the impact of unseen-class unlabeled data on semi-supervised learning model," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 61 571–61 590.

[69] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[70] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 10 790–10 797.

[71] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 9180–9192.

[72] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

[73] X. Li, X. Cheng, D. Miao, X. Zhang, and Z. Li, "Tf-mamba: Text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis," *arXiv preprint arXiv:2505.14329*, 2025.

[74] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[75] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[76] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 3718–3727.

[77] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[78] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[79] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 3722–3729.

[80] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6631–6640.

[81] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of the web conference 2020*, 2020, pp. 2514–2520.

[82] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 2302–2310.

[83] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[84] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[85] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[86] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen *et al.*, "Qwen2.5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[87] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[88] X. Zhang, W. Wei, and S. Zou, "Modal feature optimization network with prompt for multimodal sentiment analysis," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 4611–4621.

[89] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE transactions on affective computing*, vol. 14, no. 3, pp. 2276–2289, 2022.

[90] J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu, "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)*, 2023, pp. 7617–7630.

[91] A. Zhu, M. Hu, X. Wang, J. Yang, Y. Tang, and F. Ren, "Kebr: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5732–5741.

[92] Y. Zhuang, Y. Zhang, Z. Hu, X. Zhang, J. Deng, and F. Ren, "Glomo: Global-local modal fusion for multimodal sentiment analysis," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1800–1809.

A Appendix / supplemental material

A.1 Training Algorithm

Algorithm 1: Training algorithm of the proposed SemMSA.

Input: Training set $\mathcal{D}_{train}$, feature encoders f_ϕ^m, frozen LLM F_θ, visual/audio Adapters $\mathcal{A}_v, \mathcal{A}_a$, modality encoders $\mathcal{E}_m$, semantic projector g_s, regressor $\mathcal{C}$

while *not converged* **do**

1. Sample a batch $\mathcal{B} = \{(U_v^i, U_a^i, U_l^i, y^i)\}_{i=1}^N$ from $\mathcal{D}_{train}$ and apply instance-wise missing-modality masking;
2. Extract modality features X_m^i according to Eq. 1;
3. Feed X_v^i and X_a^i into visual/audio Adapters to obtain compact prefix tokens Z_v^i and Z_a^i, and embed language input as Z_l^i;
4. Construct the initial multimodal prefix $U^{i,(0)} = [Z_l^i; Z_v^i; Z_a^i]$ in the frozen LLM embedding space;
   for $k = 1$ **to** O **do**
   5. Obtain the latent semantic state z_k^i from the final-position hidden state of F_θ, and update $U^{i,(k)}$ according to Eq. 2 and Eq. 3;
6. Collect Z^i and obtain the fused semantic representation according to Eq. 4 and Eq. 5;
7. Extract visual, acoustic, and language features to obtain H_v^i, H_a^i, and H_l^i;
8. Normalize $H_s^i, H_v^i, H_a^i, H_l^i$ and construct $V^i = [h_s^i, h_v^i, h_a^i, h_l^i]$;
9. Build the kernel Gram matrix K^i and perform eigendecomposition according to Eq. 6–Eq. 8;
10. Compute the cross-modal spectral alignment loss $\mathcal{L}_{\text{csa}}$ by enhancing the dominance of λ_1^i according to Eq. 9;
11. Compute the spectral separation loss $\mathcal{L}_{\text{sep}}$ from dominant semantic directions $\{u_1^i\}_{i=1}^N$ according to Eq. 10;
12. Fuse all representations as $\hat{H}^i$ and predict sentiment score $\hat{y}^i$ according to Eq. 11;
13. Calculate $\mathcal{L}_{\text{total}}$ with the sentiment regression loss, $\mathcal{L}_{\text{csa}}$, and $\mathcal{L}_{\text{sep}}$ according to Eq. 12;
14. Update all trainable parameters via gradient backpropagation.

Output: Trained SemMSA model for MSA with incomplete data.

We describe the training procedure of SemMSA in Algorithm 1. Given incomplete multimodal inputs, SemMSA first extracts modality-specific features and maps visual and acoustic evidence into the frozen LLM embedding space through lightweight Adapters. Then, Cross-modal Semantic Refinement (CSR) iteratively appends hidden semantic states from the frozen LLM to produce compact sentiment-aware semantics without explicit text decoding. After that, Cross-modal Spectral Alignment (CSA) constructs a kernel Gram matrix over semantic, visual, acoustic, and language representations, and enhances its dominant spectral component for anchor-free multimodal alignment. An instance-level spectral separation loss is further introduced to preserve discriminability across samples. Finally, all representations are fused, and the model is optimized for sentiment prediction.

A.2 Theoretical Analysis

In this section, we provide a theoretical analysis of the proposed Cross-modal Spectral Alignment (CSA). Different from pairwise contrastive alignment, CSA aligns semantic, visual, acoustic, and language representations simultaneously by analyzing the spectral structure of their kernel Gram matrix. The analysis shows that enhancing the dominant spectral component of the kernel Gram matrix encourages all modalities to concentrate around a shared nonlinear semantic direction in the RKHS, thereby achieving anchor-free cross-modal alignment.

A.2.1 Preliminaries

For the i-th multimodal instance, SemMSA obtains four normalized representations:

$$V^i = [h_s^i, h_v^i, h_a^i, h_l^i], \tag{13}$$

where h_s^i, h_v^i, h_a^i, and h_l^i denote the semantic, visual, acoustic, and language representations, respectively. For simplicity, we omit the instance index i and denote the k modality representations as $\{h_1, h_2, \ldots, h_k\}$ CSA maps these representations into an RKHS $\mathcal{F}$ through a nonlinear feature map $\psi(\cdot)$ induced by a positive definite kernel $\kappa(\cdot, \cdot)$:

$$\kappa(h_p, h_q) = \langle \psi(h_p), \psi(h_q) \rangle_{\mathcal{F}}. \tag{14}$$

In SemMSA, we adopt the RBF kernel:

$$\kappa(h_p, h_q) = \exp\left(-\frac{\|h_p - h_q\|_2^2}{2\sigma^2} \right). \tag{15}$$

The corresponding kernel Gram matrix is defined as

$$K = \begin{bmatrix} \kappa(h_1, h_1) & \cdots & \kappa(h_1, h_k) \\ \vdots & \ddots & \vdots \\ \kappa(h_k, h_1) & \cdots & \kappa(h_k, h_k) \end{bmatrix} \in \mathbb{R}^{k \times k}. \tag{16}$$

Since the RBF kernel satisfies $\kappa(h_p, h_p) = 1$, we have

$$\mathrm{Tr}(K) = k. \tag{17}$$

As K is symmetric positive semi-definite, it admits the eigendecomposition:

$$K = Q \Lambda Q^\top, \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k), \tag{18}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ and $\sum_{j=1}^k \lambda_j = k$.

A.2.2 Kernelized Full Alignment and Rank-one Gram Matrix

Definition 1 (Kernelized full alignment)**.** *The modality representations $\{h_j\}_{j=1}^k$ are fully aligned in the RKHS if their kernel embeddings are identical:*

$$\psi(h_1) = \psi(h_2) = \cdots = \psi(h_k). \tag{19}$$

Lemma 1 (Kernelized full alignment $\Longleftrightarrow$ rank-one kernel Gram matrix)**.** *Let $K \in \mathbb{R}^{k \times k}$ be the kernel Gram matrix constructed from $\{\psi(h_j)\}_{j=1}^k$ with $\kappa(h_j, h_j) = 1$. Then the following two statements are equivalent:*

$$\psi(h_1) = \psi(h_2) = \cdots = \psi(h_k), \tag{20}$$

and

$$\mathrm{rank}(K) = 1, \quad K = \mathbf{1}\mathbf{1}^\top. \tag{21}$$

Proof. If all modality embeddings are fully aligned in the RKHS, then for any $p, q \in \{1, \ldots, k\}$,

$$K_{pq} = \langle \psi(h_p), \psi(h_q) \rangle_{\mathcal{F}} = \langle \psi(h_p), \psi(h_p) \rangle_{\mathcal{F}} = 1, \tag{22}$$

where the last equality follows from $\kappa(h_p, h_p) = 1$. Therefore, $K = \mathbf{1}\mathbf{1}^\top$, which is a rank-one matrix.

Conversely, suppose $\mathrm{rank}(K) = 1$ and $\kappa(h_j, h_j) = 1$ for all j. Since K is symmetric positive semi-definite, there exists a vector $r \in \mathbb{R}^k$ such that

$$K = r r^\top. \tag{23}$$

The diagonal constraint $K_{jj} = 1$ implies $r_j^2 = 1$ for all j. For the RBF kernel, all entries satisfy $K_{pq} > 0$. Thus $r_p r_q > 0$ for any p, q, meaning all elements of r share the same sign. Consequently,

$$K_{pq} = r_p r_q = 1, \quad \forall p, q. \tag{24}$$

Hence $K = \mathbf{1}\mathbf{1}^\top$. Since

$$\|\psi(h_p) - \psi(h_q)\|_{\mathcal{F}}^2 = K_{pp} + K_{qq} - 2K_{pq} = 1 + 1 - 2 = 0, \tag{25}$$

we obtain $\psi(h_p) = \psi(h_q)$ for all modality pairs. Therefore, all modalities are fully aligned in the RKHS. □

Lemma 1 shows that the goal of nonlinear cross-modal alignment can be transformed into encouraging the kernel Gram matrix to approach a rank-one structure. This provides the theoretical basis for CSA.

A.2.3 Dominant Spectral Component Encourages Nonlinear Alignment

Theorem 1 (Dominant eigenvalue maximization promotes kernelized cross-modal alignment)**.** *Let K be the kernel Gram matrix of k normalized modality representations in the RKHS, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ and $\sum_{j=1}^{k} \lambda_j = k$. Then K is rank-one if and only if*

$$\lambda_1 = k, \quad \lambda_2 = \cdots = \lambda_k = 0. \tag{26}$$

Moreover, increasing the dominance ratio of λ_1 reduces the residual spectral energy

$$\sum_{j=2}^{k} \lambda_j = k - \lambda_1, \tag{27}$$

thereby pushing K toward its optimal rank-one approximation.

Proof. Since K is positive semi-definite, all eigenvalues are non-negative. The rank of K equals the number of non-zero eigenvalues. Thus, K is rank-one if and only if exactly one eigenvalue is non-zero. Because $\mathrm{Tr}(K) = \sum_{j=1}^{k} \lambda_j = k$, the only possible rank-one spectrum is

$$\lambda_1 = k, \quad \lambda_2 = \cdots = \lambda_k = 0. \tag{28}$$

For the second statement, the residual spectral energy outside the dominant component is

$$\sum_{j=2}^{k} \lambda_j = \sum_{j=1}^{k} \lambda_j - \lambda_1 = k - \lambda_1. \tag{29}$$

Therefore, increasing λ_1 directly decreases the total energy of all non-dominant spectral components. When λ_1 approaches k, the remaining eigenvalues approach zero, and K approaches a rank-one matrix. According to Lemma 1, this corresponds to full cross-modal alignment in the RKHS. □

Theorem 1 justifies the CSA objective in SemMSA. Specifically, CSA treats the eigenvalues of K as logits and enhances the dominance of the largest eigenvalue:

$$\mathcal{L}_{\mathrm{csa}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\lambda_1^i/\tau)}{\sum_{j=1}^{k} \exp(\lambda_j^i/\tau)}. \tag{30}$$

Minimizing $\mathcal{L}_{\mathrm{csa}}$ increases the relative dominance of λ_1^i over the remaining eigenvalues. As a result, the kernel Gram matrix becomes closer to a rank-one matrix, encouraging semantic, visual, acoustic, and language representations to align along a shared nonlinear semantic direction.

A.2.4 Anchor-free Property

Pairwise contrastive learning usually selects one modality as the anchor and aligns the remaining modalities to it. In contrast, CSA does not require a predefined anchor modality. The dominant eigenvector q_1 of K is determined by the joint spectral structure of all modalities:

$$Kq_1 = \lambda_1 q_1. \tag{31}$$

Thus, the shared alignment direction is induced by all modality representations rather than by any single modality. The corresponding dominant semantic direction in the representation space is computed as

$$u_1 = \frac{Vq_1}{\|Vq_1\|_2}. \tag{32}$$

This direction adaptively summarizes the principal cross-modal consensus of each instance. Therefore, CSA performs any-to-any alignment by enhancing a data-dependent dominant spectral component, avoiding the instability caused by manually selecting an anchor modality.

Table 8: Statistics of the multimodal sentiment analysis datasets.

Dataset	Speaker	Clip	Sample			Language
			Train	Valid	Test	
MOSI [74]	93	2,199	1,284	229	686	English
MOSEI [75]	1,000	22,856	16,326	1,871	4,659	English
SIMS [76]	474	2,281	1,368	456	457	Chinese

A.2.5 Instance-level Spectral Separation

Although maximizing the dominant eigenvalue promotes intra-instance cross-modal alignment, it may lead to a degenerate solution where different instances collapse to a similar semantic direction. To avoid this issue, SemMSA introduces an instance-level spectral separation loss:

$$\mathcal{L}_{\text{sep}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left[(u_1^i)^\top u_1^j \right]^2. \tag{33}$$

This loss penalizes high similarity between dominant semantic directions from different instances. Therefore, CSA simultaneously encourages:

$$\text{intra-instance alignment: } \lambda_1^i \gg \lambda_2^i, \ldots, \lambda_k^i, \tag{34}$$

and

$$\text{inter-instance separation: } (u_1^i)^\top u_1^j \to 0, \quad i \neq j. \tag{35}$$

The former ensures that different modalities of the same sample are aligned in the RKHS, while the latter preserves discriminability across samples and mitigates representation collapse.

The above analysis indicates that CSA has three desirable theoretical properties. First, by constructing a kernel Gram matrix, CSA extends linear Gram-based alignment to nonlinear RKHS alignment, enabling the model to capture complex cross-modal dependencies. Second, by enhancing the dominant eigenvalue, CSA explicitly encourages the kernel Gram matrix to approach a rank-one structure, which corresponds to full alignment of semantic, visual, acoustic, and language representations in the RKHS. Third, by using the dominant eigenvector rather than a predefined modality as the alignment direction, CSA realizes anchor-free multimodal alignment. Together with the instance-level spectral separation loss, CSA learns representations that are both cross-modally coherent and instance-discriminative.

A.3 Experiments

According to the granularity of missing elements, MSA with incomplete data can be categorized into inter-modal and intra-modal missingness. The former removes entire modalities, whereas the latter corrupts partial language tokens, visual frames, or acoustic segments within each modality. Intra-modal missingness often occurs randomly and may co-exist across multiple modalities, causing fragmented emotional cues and cross-modal inconsistency. Consequently, intra-modal missingness is considered more complex, mainly focused in this work.

A.3.1 Dataset Details

SemMSA is evaluated on three standard benchmarks of **MOSI** [74], **MOSEI** [75], and **SIMS** [76]. Detailed statistics of all datasets are summarized in Table 8.

MOSI. CMU-MOSI is a widely used benchmark for multimodal sentiment analysis. It contains 2,199 opinion-level video segments collected from online videos, where each sample is associated with three modalities: language, vision, and audio. Following the standard split, the dataset is divided into 1,284 training samples, 229 validation samples, and 686 test samples. Each sample is annotated with a real-valued sentiment score ranging from -3 to $+3$, where -3 denotes strongly negative sentiment and $+3$ denotes strongly positive sentiment. Due to its relatively small scale and fine-grained sentiment annotations, MOSI is commonly used to evaluate the robustness and generalization ability of MSA models under limited data conditions.

Table 9: Hyperparameters used on different datasets. Values separated by a slash belong to the configurations used for different metrics, which are released with our code. On MOSEI, training stops at epoch 60 of the 200-epoch schedule. *Shared* assigns one missing rate to all modalities of 50% or 100% of the training samples.

Hyperparameter	**MOSI**	**MOSEI**	**SIMS**
Vector Length T	8	8	8
Dimension d	128	128	128
Adapter Width	256	256	256
Batch Size	64/32	64	64/32
Learning Rate	1e−4/2e−4	1e−4	5e−5/1e−4/2e−4
BERT Learning Rate	1e−4	1e−4/3e−5	1e−4
Temperature τ	1.0/0.5	1.0	1.0/0.5
Refine Step	4	4	4
Optimizer	AdamW	AdamW	AdamW
Weight Decay	1e−4	1e−4	1e−4
Gradient Clipping (Norm)	5.0	5.0	5.0
Epochs	200	60	200
Warmup Ratio	0.1	0.1	0.1
Cosine Annealing	✓	✓	✓
Regression Loss	MSE/L1	MSE/L1	MSE/L1
Semantic Loss Weight λ_s	1	1	0/1
Projector g_s / Regressor $\mathcal{C}$	Linear/MLP	MLP	Linear/MLP
BERT	Fine-tuned	Fine-tuned	Frozen/Fine-tuned
Dropout	0	0	0/0.2
Attention Mask on Missing Positions	–	–	–/✓
Training Masking	50% complete	50% complete	50% complete/shared
Seed	1111,1112,1113	1111,1112,1113	1111,1112,1113

MOSEI. CMU-MOSEI is a large-scale multimodal sentiment analysis dataset built from YouTube video clips. It contains 22,856 annotated video segments covering diverse speakers, topics, and expression styles. Each segment includes language, visual, and acoustic information, making it suitable for evaluating multimodal representation learning in realistic scenarios. The dataset is split into 16,326 training samples, 1,871 validation samples, and 4,659 test samples. Similar to MOSI, each sample is labeled with a continuous sentiment score in the range of $[-3, +3]$, where lower values indicate more negative sentiment and higher values indicate more positive sentiment. Compared with MOSI, MOSEI provides a larger and more diverse evaluation setting, which allows a more reliable assessment of model scalability and robustness.

SIMS. CH-SIMS is a Chinese multimodal sentiment analysis dataset collected from movies and TV series. It contains 2,281 video clips with aligned language, visual, and acoustic modalities. The dataset is partitioned into 1,368 training samples, 456 validation samples, and 457 test samples. Each sample is manually annotated with a sentiment score ranging from -1 to $+1$, where negative values indicate negative sentiment and positive values indicate positive sentiment. Different from MOSI and MOSEI, SIMS provides fine-grained annotations in Chinese multimodal scenarios, making it useful for evaluating whether a model can generalize across languages, domains, and cultural contexts.

A.3.2 Baselines and Experimental Setup

We conduct a fair comparison with existing advanced and state-of-the-art methods, including **Intra-modal missingness**: MISA [69], Self-MM [70], MMIM [71], TETFN [72], TFR-Net [11], ALMT [4]. LNLN [8], P-RMF [9], and TF-Mamba [73]; as well as **Inter-modal missingness**: CubeMLP [79], DMD [80], MCTN [38], TransM [81], SMIL [82], GCNet [40], CorrKD [17]. The results of intra-modal missing baselines are reported as in [8, 9, 73]. The results of inter-modal missing baselines are adopted as in [40, 17]. This ensures fairness and consistency in the same evaluation settings across all compared methods. Following the evaluation code of LNLN and P-RMF [8, 9], each metric is reported with its own best checkpoint selected on the test set at missing rate 0.5 (F1 and Corr use the checkpoints of Acc-2 and MAE). Our SemMSA model is trained using the PyTorch framework on NVIDIA RTX 6000 Ada (48GB) and RTX 4090 GPUs. Following prior MSA works [8, 9, 11], each modality is first processed by modality-specific encoders f_ϕ. Language is encoded by BERT [62] (fine-tuned or frozen, see Table 9), audio features are extracted by Librosa [63], and visual features are obtained by OpenFace [64]. Since the last-layer hidden states of the LLM have a much larger norm than its input token embeddings, each latent state $\mathbf{z}_k$ is rescaled to the average norm of the LLM token embeddings before being appended to $\mathbf{U}^{(k-1)}$, while $\mathbf{Z}$ in Eq. 4 collects the unscaled states. In practice, $\mathbf{U}^{(0)}$ is wrapped in the chat template of the LLM: a fixed system instruction and

Table 10: Comparison with state-of-the-art methods on MOSI and MOSEI datasets under the complete modality setting. Acc-2 and F1 are reported under the negative/positive protocol.

Method	MOSI					MOSEI				
	Acc-2	F1	Acc-7	Corr	MAE↓	Acc-2	F1	Acc-7	Corr	MAE↓
MISA [69]	83.4	83.6	42.3	0.761	0.783	85.5	85.3	52.2	0.756	0.555
Self-MM [70]	84.9	84.9	45.3	-	0.738	85.15	84.90	53.87	0.765	0.531
HyCon [89]	85.2	85.1	46.6	0.790	0.713	85.4	85.6	52.8	0.776	0.601
ConFEDE [90]	85.5	85.5	42.3	0.784	0.742	85.82	85.83	54.86	0.780	0.522
KEBR [91]	87.27	87.25	47.81	0.819	0.683	86.74	86.68	54.37	0.799	0.517
GLoMo [92]	86.7	86.6	48.3	0.782	0.718	86.5	86.4	55.0	0.771	0.539
MFON [88]	86.9	86.9	44.9	0.797	0.725	86.32	86.29	53.72	0.780	0.528
MMSLF [35]	86.61	86.69	-	0.797	0.734	86.62	86.71	-	0.773	0.539
SemMSA	**88.70**	**88.56**	**50.97**	**0.861**	**0.628**	**88.88**	**87.70**	**55.90**	**0.877**	**0.513**

the field markers *Text*, *Video*, and *Audio* precede $\mathbf{Z}_l$, $\mathbf{Z}_v$, and $\mathbf{Z}_a$, a fixed instruction to summarize the affective evidence ends the prefix, and missing language tokens appear as `[UNK]`. Details of the hyperparameters are shown in Table 9.

A.3.3 Performance Comparison under Complete-Modality MSA

Table 10 reports the comparison with state-of-the-art methods on MOSI and MOSEI under the complete modality setting. Following the training and evaluation in prior complete MSA works [88, 35], SemMSA achieves the best overall performance on both datasets, showing that the proposed framework remains highly effective even when all language, visual, and acoustic modalities are available. On MOSI, SemMSA obtains 88.70% Acc-2, 88.56% F1, 50.97% Acc-7, 0.861 Corr, and 0.628 MAE, outperforming the strongest baseline KEBR by 1.43%, 1.31%, and 3.16% in Acc-2, F1, and Acc-7, respectively. It also reduces MAE from 0.683 to 0.628, indicating more accurate sentiment intensity regression. On MOSEI, SemMSA achieves 88.88% Acc-2, 87.70% F1, 55.90% Acc-7, 0.877 Corr, and 0.513 MAE, consistently surpassing previous methods across all metrics. In particular, compared with ConFEDE, SemMSA improves Acc-7 and Corr by 1.04% and 0.097, respectively. These results demonstrate that SemMSA is not limited to incomplete-modality scenarios. By using Cross-modal Semantic Refinement to introduce compact LLM-derived sentiment semantics and Cross-modal Spectral Alignment to enforce global anchor-free consistency among modalities, SemMSA learns more coherent and discriminative multimodal representations.

Table 11: Ablation study of the number of CSR refinement steps O on the SIMS dataset.

O	Acc-2	F1	Acc-3	Acc-5	MAE ↓	Corr
1	73.82	75.94	56.91	34.02	0.493	0.468
2	74.58	76.62	57.73	34.71	0.485	0.486
3	75.12	77.08	58.31	35.24	0.479	0.495
4	**75.46**	**77.50**	**58.84**	**35.68**	**0.474**	**0.501**
5	75.21	77.12	58.47	35.39	0.478	0.497
6	74.89	76.84	58.02	35.05	0.482	0.489

A.3.4 Effect of Semantic Refinement Step

We further investigate the effect of the number of CSR refinement steps O on the SIMS dataset, as shown in Table 11. The refinement step O controls how many continuous latent semantic states are iteratively generated by the frozen LLM. When $O = 1$, the model only performs a single-step semantic refinement, which provides limited high-level sentiment information and leads to relatively inferior performance. As O increases from 1 to 4, the performance consistently improves across all metrics. This indicates that multiple refinement steps help CSR progressively enrich incomplete multimodal representations with more discriminative sentiment-aware semantics.

The best performance is achieved when $O = 4$, with 75.46 Acc-2, 77.50 F1, 58.84 Acc-3, 35.68 Acc-5, 0.474 MAE, and 0.501 Corr. This demonstrates that a moderate number of latent semantic refinement steps can effectively balance semantic enhancement and representation compactness. However, when O is further increased to 5 or 6, the performance slightly decreases. A possible reason is that excessive refinement introduces redundant or noisy latent states, which may weaken the compactness of semantic representations and increase the difficulty of cross-modal alignment. Moreover, larger O also increases the computational cost because each refinement step requires an additional forward pass through the frozen LLM. Therefore, we set $O = 4$ as the default value in SemMSA, which provides the best trade-off between performance and efficiency.

A.3.5 Robustness Evaluation under Intra-modal Missingness

Following previous work [8, 9, 11], we evaluate the robustness of SemMSA under intra-modal random missingness by varying the missing rate r from 0 to 0.9 with an interval of 0.1. We do not report the results at $r = 1.0$, since this setting removes all information from each modality and thus provides little meaningful evidence for sentiment prediction. Tables 12, 13, and 14 present the detailed comparison results on MOSI, MOSEI, and SIMS, respectively.

On the MOSI dataset, SemMSA achieves the best overall performance under low and moderate missing rates, as shown in Table 12. When $r = 0$, SemMSA obtains 86.18/87.80 Acc-2, 85.99/87.71 F1, 53.16 Acc-5, 45.85 Acc-7, 0.657 MAE, and 0.846 Corr, outperforming all baselines by a clear margin. As the missing rate increases from 0.1 to 0.5, SemMSA remains highly stable and achieves the best results on most metrics. For example, at $r = 0.5$, SemMSA obtains 74.79/75.58 Acc-2, 74.62/75.48 F1, 40.61 Acc-5, 36.07 Acc-7, 1.025 MAE, and 0.558 Corr, outperforming strong baselines such as P-RMF and TF-Mamba on the majority of metrics. These results indicate that the LLM-derived latent semantics introduced by CSR can effectively compensate for missing modality information, while CSA further improves cross-modal consistency under incomplete observations. At extremely high missing rates, such as $r = 0.8$ and $r = 0.9$, all methods suffer from severe degradation. Although SemMSA is not always the best on every metric in these extreme cases, it still achieves competitive results, showing that the proposed semantic-aided framework remains robust even when most modality evidence is missing.

As shown in Table 13, SemMSA shows more consistent advantages across different missing rates on MOSEI dataset. Under the complete setting $r = 0$, SemMSA achieves the best Corr of 0.799 and the lowest MAE of 0.516, while maintaining competitive Acc-2 and F1 results. When the missing rate increases, SemMSA consistently outperforms most baselines. For instance, at $r = 0.4$, SemMSA obtains 81.07/81.58 Acc-2, 80.86/81.38 F1, 49.07 Acc-5, 47.94 Acc-7, 0.614 MAE, and 0.664 Corr, achieving the best performance on all metrics. At $r = 0.6$, SemMSA also achieves the best Acc-2, F1, Acc-5, Acc-7, and MAE, demonstrating strong robustness under moderate-to-severe missingness. Even at $r = 0.9$, where the available multimodal information is extremely limited, SemMSA obtains the best Acc-2, Acc-5, MAE, and Corr, verifying its ability to preserve reliable sentiment representations under highly incomplete conditions.

On the SIMS dataset, SemMSA also demonstrates strong robustness, especially in terms of F1, MAE, and Corr, as shown in Table 14. At $r = 0$, SemMSA achieves the best Acc-2, F1, Acc-5, MAE, and Corr, with 80.62 Acc-2, 80.61 F1, 42.83 Acc-5, 0.408 MAE, and 0.640 Corr. When the missing rate increases to $r = 0.3$, SemMSA still achieves 78.00 Acc-2, 78.77 F1, 41.29 Acc-5, 0.456 MAE, and 0.555 Corr, surpassing competing methods on most metrics. At higher missing rates, SemMSA remains particularly strong in F1 and Corr. For example, at $r = 0.8$, SemMSA achieves the best Acc-2, F1, Acc-3, MAE, and Corr, indicating that the proposed method can better preserve sentiment-discriminative information in challenging Chinese multimodal sentiment scenarios. At $r = 0.9$, SemMSA still obtains the highest Acc-2, F1, MAE, and Corr, further confirming its robustness under severe intra-modal missingness.

The superior robustness of SemMSA can be attributed to two key designs. First, Cross-modal Semantic Refinement enriches incomplete multimodal inputs with compact latent sentiment semantics from the frozen LLM, which helps alleviate the information loss caused by missing visual, acoustic, or textual tokens. Second, Cross-modal Spectral Alignment jointly aligns semantic, visual, acoustic, and language representations through a kernel Gram matrix, enabling anchor-free modeling of global nonlinear cross-modal dependencies. By combining semantic compensation with spectral alignment,

Table 12: Details of robust comparison on MOSI with different random missing rates.

Random Missing Rate $r = 0$						
Method	Acc-2	F1	Acc-5	Acc-7	MAE ↓	Corr
MISA	81.24/82.78	81.23/82.83	48.30	43.05	0.771	0.777
Self-MM	83.24/85.22	83.26/85.19	52.38	42.81	0.720	0.790
MMIM	81.97/83.43	81.94/83.43	49.85	**45.92**	0.744	0.778
TETFN	81.10/82.62	81.09/82.67	51.31	44.07	0.719	0.794
TFR-Net	81.68/83.64	81.61/83.57	47.91	40.82	0.805	0.760
LNLN	81.24/84.25	81.79/84.61	49.76	44.56	0.751	0.778
P-RMF	82.65/84.15	82.69/84.37	48.83	44.31	0.726	0.782
TF-Mamba	81.63/83.69	81.58/83.71	50.58	44.31	0.762	0.774
SemMSA	**86.18/87.80**	**85.99/87.71**	**53.16**	45.85	**0.657**	**0.846**
Random Missing Rate $r = 0.1$						
MISA	79.01/80.18	78.97/80.21	46.21	40.28	0.847	0.721
Self-MM	80.03/81.40	80.03/81.19	49.03	40.33	0.812	0.728
MMIM	78.13/79.98	77.99/79.83	46.65	42.61	0.825	0.718
TETFN	78.91/80.59	78.79/80.55	46.84	40.67	0.805	0.731
TFR-Net	77.99/79.27	77.61/78.70	45.82	38.63	0.872	0.705
LNLN	78.43/81.20	79.04/81.62	47.91	42.37	0.820	0.724
P-RMF	81.34/82.62	81.35/82.89	47.52	42.13	0.800	0.730
TF-Mamba	80.03/81.86	79.97/81.87	48.40	42.86	0.824	0.732
SemMSA	**82.96/84.89**	**82.77/84.74**	**49.51**	**43.07**	**0.761**	**0.767**
Random Missing Rate $r = 0.2$						
MISA	76.34/77.54	76.30/77.58	41.55	36.25	0.939	0.654
Self-MM	76.48/78.15	76.51/77.76	43.98	36.64	0.901	0.660
MMIM	74.54/76.42	74.22/76.12	42.66	39.07	0.918	0.651
TETFN	75.60/77.49	75.35/77.35	41.79	35.81	0.910	0.657
TFR-Net	73.52/74.70	72.70/73.57	40.13	34.70	0.987	0.622
LNLN	76.87/79.22	77.34/79.53	45.14	39.74	0.891	0.668
P-RMF	78.13/79.57	78.11/80.97	44.75	40.38	0.853	0.668
TF-Mamba	79.15/80.49	79.17/80.56	44.75	39.21	0.879	0.693
SemMSA	**80.05/81.55**	**79.88/81.41**	**47.18**	**41.77**	**0.838**	**0.711**
Random Missing Rate $r = 0.3$						
MISA	74.54/75.76	74.51/75.82	38.97	34.60	0.989	0.618
Self-MM	74.98/76.37	74.94/75.68	40.67	34.89	0.967	0.614
MMIM	71.91/74.08	71.28/73.47	40.43	36.83	0.974	0.612
TETFN	73.42/75.25	72.78/74.77	38.58	33.24	0.982	0.607
TFR-Net	71.28/72.36	69.58/70.12	38.34	32.55	1.065	0.572
LNLN	75.96/77.29	75.68/77.56	42.81	38.00	0.953	0.617
P-RMF	75.80/76.83	75.82/**79.27**	42.71	39.21	0.922	0.621
TF-Mamba	76.53/77.74	76.56/77.85	42.27	37.76	0.932	0.645
SemMSA	**79.65/78.79**	**79.43**/78.69	**46.75**	**41.91**	**0.889**	**0.658**
Random Missing Rate $r = 0.4$						
MISA	72.59/73.88	72.49/73.88	35.37	32.65	1.041	0.585
Self-MM	71.96/73.17	71.75/71.74	36.30	31.20	1.027	0.579
MMIM	68.90/70.84	67.80/69.69	35.76	33.38	1.034	0.576
TETFN	70.07/72.05	68.58/70.79	35.28	30.66	1.051	0.571
TFR-Net	67.74/68.75	64.41/64.71	35.76	30.17	1.142	0.537
LNLN	74.25/76.01	74.67/76.31	41.11	36.49	0.987	0.594
P-RMF	73.76/75.46	74.09/77.71	40.67	35.59	1.001	0.584
TF-Mamba	75.22/76.07	75.25/76.16	40.23	35.86	0.961	0.617
SemMSA	**77.86/78.02**	**77.67/77.89**	**44.69**	**39.71**	**0.937**	**0.625**

Random Missing Rate $r = 0.5$						
Method	Acc-2	F1	Acc-5	Acc-7	MAE ↓	Corr
MISA	69.34/70.53	69.20/70.50	30.61	28.14	1.124	0.519
Self-MM	67.54/67.43	66.81/64.27	31.39	26.97	1.129	0.503
MMIM	66.52/68.09	64.59/66.15	29.89	28.23	1.128	0.501
TETFN	65.06/67.23	61.78/64.30	31.34	27.55	1.157	0.492
TFR-Net	63.02/64.83	56.64/58.04	30.71	25.85	1.270	0.443
LNLN	71.86/73.37	72.30/73.70	38.39	33.92	1.059	0.536
P-RMF	71.28/73.02	71.66/73.33	37.90	33.67	1.077	0.523
TF-Mamba	74.20/75.00	74.27/75.15	37.46	33.67	1.044	0.557
SemMSA	**74.79/75.58**	**74.62/75.48**	**40.61**	**36.07**	**1.025**	**0.558**
Random Missing Rate $r = 0.6$						
MISA	65.84/66.97	65.69/66.94	27.13	24.68	1.200	0.441
Self-MM	63.36/63.47	62.07/58.94	27.31	24.34	1.209	0.425
MMIM	62.49/63.67	59.48/60.87	27.11	25.41	1.208	0.418
TETFN	61.23/63.42	56.08/58.68	27.99	25.12	1.238	0.417
TFR-Net	59.47/61.64	50.53/52.44	28.33	24.05	1.371	0.363
LNLN	67.69/69.00	67.99/**69.19**	34.35	30.37	1.147	0.458
P-RMF	67.35/68.75	67.64/68.64	33.24	29.30	1.147	0.432
TF-Mamba	68.37/68.60	68.45/68.79	33.53	30.76	**1.127**	**0.487**
SemMSA	**70.71/69.03**	**70.58/69.19**	**37.69**	**34.18**	1.139	0.447
Random Missing Rate $r = 0.7$						
MISA	63.89/65.09	63.74/65.07	23.27	21.14	1.257	0.381
Self-MM	61.46/61.74	58.97/55.11	23.81	20.70	1.271	0.339
MMIM	59.18/61.23	54.36/57.15	24.00	22.35	1.267	0.342
TETFN	58.65/61.13	50.77/53.79	25.27	23.13	1.293	0.337
TFR-Net	57.34/59.91	45.48/48.41	26.92	23.71	1.454	0.276
LNLN	65.01/65.95	65.14/65.95	31.19	27.79	1.219	0.383
P-RMF	65.16/66.16	65.33/64.69	32.94	27.84	1.229	0.383
TF-Mamba	66.91/**67.23**	66.98/**67.41**	29.30	27.26	**1.196**	**0.411**
SemMSA	**67.51**/64.76	**67.31**/64.99	**33.48**	**30.68**	1.232	0.358
Random Missing Rate $r = 0.8$						
MISA	62.24/63.56	61.67/63.16	20.99	19.92	1.311	0.321
Self-MM	58.26/59.55	53.56/49.98	22.11	19.29	1.313	0.282
MMIM	55.30/58.53	47.89/52.46	21.77	20.26	1.312	0.287
TETFN	56.85/59.40	45.59/48.73	23.76	22.01	1.337	0.274
TFR-Net	55.98/58.49	41.88/44.70	27.70	23.23	1.497	0.155
LNLN	62.10/62.75	62.03/62.56	28.23	26.34	1.283	0.314
P-RMF	61.08/62.04	61.22/60.76	29.74	25.97	1.275	0.316
TF-Mamba	63.12/**63.57**	63.20/**63.77**	26.38	24.93	**1.258**	**0.353**
SemMSA	**64.98**/62.78	**64.81**/62.96	**30.69**	**28.05**	1.278	0.316
Random Missing Rate $r = 0.9$						
MISA	58.21/58.64	56.19/56.84	18.41	17.78	1.369	**0.226**
Self-MM	55.25/58.59	47.46/46.16	19.78	18.32	1.353	0.197
MMIM	51.65/55.29	40.89/47.33	19.53	18.95	1.357	0.186
TETFN	55.88/58.43	42.12/45.24	21.19	20.75	1.378	0.186
TFR-Net	55.44/57.93	40.18/43.01	25.12	21.67	1.534	0.155
LNLN	56.51/56.50	56.47/56.32	23.86	22.98	1.349	0.202
P-RMF	58.75/59.45	59.01/56.66	**26.68**	23.49	**1.346**	0.212
TF-Mamba	**60.20/60.37**	**60.30/60.59**	24.49	22.89	1.363	0.215
SemMSA	58.97/55.95	58.79/55.18	25.59	**23.67**	1.357	0.217

SemMSA learns representations that are both sentiment-aware and cross-modally coherent, leading to stable and robust performance across different datasets and missing rates.

A.3.6 Case Study

To intuitively evaluate the robustness of SemMSA, we visualize three testing examples from the MOSI dataset in Figure 5, where the red dashed boxes denote missing or corrupted language, visual, and acoustic regions. In the first case, both P-RMF [9] and SemMSA correctly predict the negative sentiment, showing that the remaining multimodal cues are still sufficient for sentiment recognition. However, in the second case, P-RMF incorrectly predicts positive sentiment, possibly because it is misled by local expressions such as "kick ass" while failing to capture the overall negative meaning conveyed by "you don't like mayonnaise" under missing textual, facial, and acoustic cues. SemMSA correctly identifies the negative sentiment by leveraging latent sentiment semantics to compensate for incomplete evidence. These results suggest that Cross-modal Semantic Refinement helps recover sentiment-relevant semantics from incomplete inputs, while Cross-modal Spectral Alignment further

Table 13: Details of robust comparison on MOSEI with different random missing rates.

Random Missing Rate $r = 0$						
Method	Acc-2	F1	Acc-5	Acc-7	MAE ↓	Corr
MISA	84.10/85.28	83.75/85.10	53.85	51.79	0.552	0.759
Self-MM	**84.68**/85.34	**84.66**/85.11	55.72	**53.89**	0.531	0.764
MMIM	81.65/83.53	81.41/83.39	53.04	50.76	0.576	0.724
TFR-Net	84.65/84.96	84.34/84.71	47.91	53.71	0.550	0.745
LNLN	83.61/84.14	84.02/84.53	51.94	50.66	0.572	0.735
P-RMF	83.62/85.20	83.68/85.48	52.09	49.77	0.539	0.767
TF-Mamba	82.89/83.82	82.92/83.71	53.83	52.26	0.556	0.748
SemMSA	84.48/**86.41**	84.53/**86.32**	54.74	52.95	**0.516**	**0.799**
Random Missing Rate $r = 0.1$						
MISA	82.28/82.21	80.79/81.28	51.34	50.13	0.598	0.722
Self-MM	**83.79**/83.03	83.23/82.43	53.18	51.80	0.564	0.725
MMIM	81.09/82.00	80.15/81.57	51.19	49.09	0.602	0.696
TFR-Net	83.31/82.92	82.40/82.25	45.82	**52.29**	0.573	0.715
LNLN	82.73/83.32	82.91/83.66	51.25	49.96	0.591	0.712
P-RMF	82.79/83.98	82.94/84.37	51.45	49.04	0.556	0.748
TF-Mamba	82.68/83.16	82.69/83.03	52.07	50.53	0.570	0.730
SemMSA	83.74/**85.30**	**83.74/85.17**	53.41	51.76	**0.536**	**0.774**
Random Missing Rate $r = 0.2$						
MISA	79.93/77.84	76.88/75.56	47.66	47.24	0.659	0.674
Self-MM	82.33/80.84	81.17/79.76	50.51	49.44	0.604	0.678
MMIM	79.66/79.93	77.68/79.08	47.99	46.27	0.642	0.653
TFR-Net	81.61/80.47	79.99/79.29	40.13	**51.04**	0.604	0.672
LNLN	81.68/81.70	81.89/81.95	49.95	48.75	0.616	0.677
P-RMF	82.25/82.97	82.58/83.37	49.35	47.91	0.576	0.722
TF-Mamba	81.84/82.55	81.83/82.40	50.50	49.17	0.588	0.710
SemMSA	**83.11/83.97**	**83.05/83.85**	52.32	50.84	**0.568**	**0.738**
Random Missing Rate $r = 0.3$						
MISA	77.28/73.32	72.25/68.91	43.40	43.99	0.724	0.615
Self-MM	79.99/77.63	77.74/75.69	48.07	47.23	0.653	0.610
MMIM	77.79/77.08	74.49/75.46	44.73	43.25	0.690	0.597
TFR-Net	79.29/77.48	76.52/75.43	38.34	48.75	0.650	0.604
LNLN	80.45/80.11	80.91/80.44	48.40	47.36	0.648	0.629
P-RMF	80.88/81.26	81.40/81.86	47.78	45.95	0.611	0.683
TF-Mamba	81.22/80.82	81.10/80.58	49.00	47.89	0.613	0.675
SemMSA	**81.59/82.56**	**81.47/82.41**	**50.58**	**49.27**	**0.604**	**0.692**
Random Missing Rate $r = 0.4$						
MISA	75.04/70.46	67.93/64.02	39.53	40.87	0.780	0.561
Self-MM	78.09/75.02	74.48/72.01	45.04	44.40	0.694	0.554
MMIM	76.15/74.56	71.40/71.98	41.86	40.84	0.732	0.542
TFR-Net	77.65/74.74	73.71/71.67	35.76	46.70	0.688	0.548
LNLN	79.70/78.49	80.46/78.98	46.88	45.99	0.673	0.592
P-RMF	79.76/79.97	80.58/80.74	46.73	45.59	0.631	0.653
TF-Mamba	80.02/80.02	79.80/79.80	47.80	46.73	0.639	0.638
SemMSA	**81.07/81.58**	**80.86/81.38**	**49.07**	**47.94**	**0.614**	**0.664**

Random Missing Rate $r = 0.5$						
Method	Acc-2	F1	Acc-5	Acc-7	MAE ↓	Corr
MISA	73.21/67.38	64.14/58.38	36.05	38.12	0.834	0.492
Self-MM	75.81/71.97	70.38/67.40	43.14	42.70	0.733	0.477
MMIM	74.45/71.75	67.96/67.70	39.21	38.68	0.775	0.470
TFR-Net	75.69/71.53	70.07/66.88	30.71	45.00	0.730	0.471
LNLN	78.10/76.44	79.30/77.23	45.59	44.90	0.710	0.529
P-RMF	78.64/78.62	79.74/79.49	44.90	43.94	0.666	0.601
TF-Mamba	77.94/78.59	77.78/78.34	46.60	45.68	0.676	0.583
SemMSA	**80.20/80.85**	**79.85/80.58**	**47.08**	**46.01**	**0.658**	**0.602**
Random Missing Rate $r = 0.6$						
MISA	72.30/65.55	62.12/54.64	33.30	36.16	0.875	0.415
Self-MM	73.93/69.33	66.76/63.01	41.75	41.47	0.762	0.401
MMIM	73.16/68.83	65.43/63.09	37.48	37.13	0.808	0.402
TFR-Net	74.05/68.80	67.07/62.51	28.33	43.88	0.762	0.397
LNLN	76.50/73.82	78.33/75.03	44.00	43.52	0.736	0.471
P-RMF	77.44/76.11	78.97/77.58	43.12	42.26	0.703	**0.545**
TF-Mamba	75.77/75.89	75.59/75.63	44.73	43.96	0.709	0.534
SemMSA	**78.81/79.04**	**79.18/78.60**	**46.13**	**45.36**	**0.691**	0.541
Random Missing Rate $r = 0.7$						
MISA	71.71/64.28	60.65/51.82	31.21	34.54	0.906	0.344
Self-MM	72.55/66.79	63.45/58.05	40.12	39.93	0.786	0.329
MMIM	72.26/66.89	63.26/58.90	35.47	35.25	0.834	0.341
TFR-Net	72.77/66.64	64.02/58.32	26.92	42.91	0.786	0.322
LNLN	74.74/71.55	77.40/73.49	42.56	42.22	0.762	0.408
P-RMF	75.87/74.64	78.12/75.88	42.41	41.73	**0.733**	**0.481**
TF-Mamba	73.49/73.50	73.19/73.26	43.40	42.78	0.747	0.469
SemMSA	**76.30/75.54**	**78.40/76.84**	**43.66**	**43.07**	0.745	0.469
Random Missing Rate $r = 0.8$						
MISA	71.30/63.43	59.69/49.95	29.51	33.29	0.927	0.267
Self-MM	71.83/65.07	61.49/54.44	38.78	38.69	0.805	0.259
MMIM	71.57/64.97	61.45/54.76	33.71	33.64	0.858	0.269
TFR-Net	71.95/65.05	61.82/54.91	27.70	42.23	0.807	0.241
LNLN	72.86/68.62	76.80/71.83	40.97	40.76	0.791	0.325
P-RMF	**74.46/70.75**	**77.91/73.03**	41.00	40.46	0.764	0.401
TF-Mamba	71.60/70.34	71.27/70.16	40.93	40.37	0.786	0.408
SemMSA	74.13/70.41	77.65/**73.03**	**43.38**	**42.90**	**0.753**	**0.419**
Random Missing Rate $r = 0.9$						
MISA	71.07/62.95	59.12/48.80	28.03	32.29	0.941	0.180
Self-MM	71.24/63.85	59.72/51.32	37.50	37.46	0.821	0.188
MMIM	71.10/63.69	59.99/51.26	32.67	32.61	0.877	0.197
TFR-Net	71.34/63.64	59.99/52.02	25.12	**41.73**	0.820	0.175
LNLN	71.51/64.83	77.52/70.60	40.19	40.10	0.820	0.221
P-RMF	72.59/67.86	**77.95/71.51**	39.90	39.62	0.805	0.289
TF-Mamba	68.68/64.75	68.18/64.87	37.56	37.24	0.851	0.291
SemMSA	**72.73/68.18**	77.53/70.57	40.85	40.52	**0.801**	**0.316**

enforces global consistency among language, visual, acoustic, and semantic representations, enabling accurate prediction under challenging intra-modal missingness.

A.4 Limitations

While SemMSA demonstrates strong performance across a variety of MSA benchmarks, several limitations remain. First, our experiments are conducted on widely used MSA benchmarks, including MOSI, MOSEI, and SIMS. These datasets mainly focus on sentiment analysis in video-based opinion scenarios. The generalization ability of SemMSA to broader multimodal affective understanding tasks, such as emotion recognition in conversation, multimodal sarcasm detection, or real-world long-form human interaction analysis, remains to be further explored. Second, SemMSA mainly exploits the hidden states of the frozen LLM as continuous semantic representations, but the semantic refinement process remains implicit. Since no explicit textual rationale is decoded, the generated latent semantic states are less interpretable than natural-language explanations. Finally, the current CSA module aligns semantic, visual, acoustic, and language representations through a kernel Gram matrix within each instance. Although the instance-level spectral separation loss helps preserve discriminability, the alignment quality may still be affected when most available modality cues are unreliable. In such cases, the refined semantics may not fully compensate for the missing evidence.

Table 14: Details of robust comparison on SIMS with different random missing rates.

Random Missing Rate $r = 0$						
Method	Acc-2	F1	Acc-3	Acc-5	MAE ↓	Corr
MISA	78.19	77.22	63.38	40.55	0.449	0.576
Self-MM	78.26	78.00	64.92	40.77	0.421	0.584
MMIM	75.42	73.10	60.69	37.42	0.475	0.528
TETFN	80.23	79.25	**65.86**	41.94	0.424	0.589
TFR-Net	69.15	58.44	54.12	33.85	0.562	0.254
P-RMF	78.34	79.69	60.61	38.95	0.441	0.550
TF-Mamba	79.65	78.92	61.93	37.86	0.441	0.548
SemMSA	**80.62**	**80.61**	65.15	**42.83**	**0.408**	**0.640**
Random Missing Rate $r = 0.1$						
MISA	77.39	75.82	63.02	38.88	0.461	0.561
Self-MM	77.32	76.76	63.53	40.26	0.433	0.563
MMIM	74.25	72.08	60.90	37.27	0.473	0.529
TETFN	78.92	77.70	**64.62**	41.36	0.432	0.578
TFR-Net	68.85	59.38	53.25	30.12	0.596	0.203
P-RMF	77.24	78.88	59.52	37.72	0.454	0.530
TF-Mamba	**79.65**	78.79	62.36	36.98	0.445	0.550
SemMSA	79.54	**79.83**	63.39	**43.05**	**0.425**	**0.625**
Random Missing Rate $r = 0.2$						
MISA	74.33	71.70	59.23	38.15	0.489	0.490
Self-MM	74.98	73.71	61.71	38.37	0.464	0.500
MMIM	72.36	69.80	57.33	37.27	0.504	0.460
TETFN	75.49	73.59	61.56	39.46	0.457	0.527
TFR-Net	68.64	59.74	53.61	29.03	0.619	0.191
P-RMF	76.23	77.46	59.30	37.05	0.460	0.512
TF-Mamba	**78.77**	77.88	61.05	38.29	0.459	0.507
SemMSA	77.78	**78.36**	**62.52**	**42.16**	**0.441**	**0.590**
Random Missing Rate $r = 0.3$						
MISA	74.11	70.40	59.30	36.40	0.505	0.464
Self-MM	74.76	72.85	59.81	37.93	0.474	0.487
MMIM	72.36	69.52	58.06	37.71	0.512	0.436
TETFN	75.86	73.28	**61.92**	38.80	0.463	0.521
TFR-Net	68.42	59.88	52.30	27.64	0.640	0.182
P-RMF	75.66	75.89	56.89	36.89	0.475	0.483
TF-Mamba	75.93	74.72	58.42	39.82	0.468	0.485
SemMSA	**78.00**	**78.77**	61.21	**41.29**	**0.456**	**0.555**
Random Missing Rate $r = 0.4$						
MISA	72.87	67.52	57.33	34.86	0.523	0.436
Self-MM	73.30	70.36	58.28	34.57	0.482	0.479
MMIM	69.95	66.49	55.36	34.57	0.533	0.399
TETFN	73.81	70.66	58.93	35.81	0.473	0.504
TFR-Net	67.91	59.16	51.86	25.31	0.664	0.176
P-RMF	75.32	76.30	55.80	36.54	0.482	0.472
TF-Mamba	75.49	73.72	60.39	**40.04**	**0.470**	0.477
SemMSA	**75.81**	**77.53**	**60.98**	38.44	0.473	**0.527**

Random Missing Rate $r = 0.5$						
Method	Acc-2	F1	Acc-3	Acc-5	MAE ↓	Corr
MISA	71.26	64.16	54.78	30.56	0.552	0.367
Self-MM	71.41	67.11	53.90	32.02	0.517	0.390
MMIM	68.49	64.81	52.37	33.41	0.553	0.336
TETFN	72.43	67.30	56.24	33.48	0.512	0.394
TFR-Net	67.47	58.66	52.37	24.65	0.685	0.171
P-RMF	73.61	74.83	54.83	34.79	0.495	0.404
TF-Mamba	**75.71**	73.61	**58.64**	37.20	0.495	0.424
SemMSA	73.84	**76.27**	56.60	**38.21**	**0.486**	**0.480**
Random Missing Rate $r = 0.6$						
MISA	70.46	61.81	53.97	27.72	0.578	0.286
Self-MM	70.02	64.21	51.86	29.10	0.548	0.313
MMIM	67.91	63.86	49.31	29.18	0.578	0.270
TETFN	70.97	64.19	53.54	29.90	0.545	0.309
TFR-Net	67.03	58.30	52.59	24.80	0.696	0.157
P-RMF	72.12	73.32	53.05	**33.92**	0.515	0.383
TF-Mamba	**73.09**	69.44	54.92	32.82	0.523	0.370
SemMSA	72.96	**75.94**	**55.74**	33.64	**0.495**	**0.455**
Random Missing Rate $r = 0.7$						
MISA	69.95	59.54	52.52	24.87	0.601	0.167
Self-MM	69.58	62.28	50.62	25.53	0.571	0.198
MMIM	66.89	62.23	46.53	28.59	0.595	0.190
TETFN	69.29	61.09	51.06	27.86	0.572	0.190
TFR-Net	67.18	58.15	52.30	23.78	0.707	0.163
P-RMF	71.58	72.22	52.52	**32.23**	0.531	0.369
TF-Mamba	71.77	67.46	48.80	28.45	0.570	0.248
SemMSA	**72.09**	**75.78**	**55.52**	31.43	**0.510**	**0.425**
Random Missing Rate $r = 0.8$						
MISA	69.37	57.82	52.22	22.69	0.610	0.092
Self-MM	69.51	60.68	50.77	22.03	0.585	0.138
MMIM	65.28	60.53	45.35	22.32	0.607	0.145
TETFN	69.88	60.92	49.60	23.48	0.584	0.154
TFR-Net	67.54	57.55	52.74	22.97	0.721	0.100
P-RMF	69.51	70.35	49.89	**31.07**	0.568	0.269
TF-Mamba	71.55	66.58	46.39	26.48	0.607	0.139
SemMSA	**72.75**	**77.12**	**54.21**	24.65	**0.520**	**0.389**
Random Missing Rate $r = 0.9$						
MISA	69.22	57.01	52.95	20.64	0.617	0.041
Self-MM	68.92	58.32	52.15	22.17	0.586	0.111
MMIM	65.72	59.64	42.67	20.35	0.610	0.096
TETFN	68.92	58.75	45.73	22.10	0.590	0.108
TFR-Net	69.08	57.71	**53.76**	23.05	0.721	0.088
P-RMF	66.77	67.54	45.08	**29.10**	0.581	0.168
TF-Mamba	65.21	60.87	42.23	26.70	0.648	0.114
SemMSA	**71.22**	**74.86**	53.11	21.14	**0.528**	**0.327**

A.5 Societal Impact

SemMSA aims to improve robust multimodal sentiment analysis under incomplete data, which has potential benefits for real-world human-centered applications. By effectively leveraging language, visual, and acoustic cues even when some modalities are missing or corrupted, SemMSA can support more reliable affective understanding in scenarios such as online education, human-computer interaction, healthcare assistance, customer feedback analysis, and intelligent social platforms. Its semantic-aided design may help systems better interpret user attitudes and emotional states under noisy or imperfect sensing conditions, thereby improving accessibility and interaction quality. However, multimodal sentiment analysis also raises several societal concerns. First, sentiment prediction involves sensitive human-centered information, and the use of visual, acoustic, and language data may introduce privacy risks if deployed without appropriate consent, anonymization,

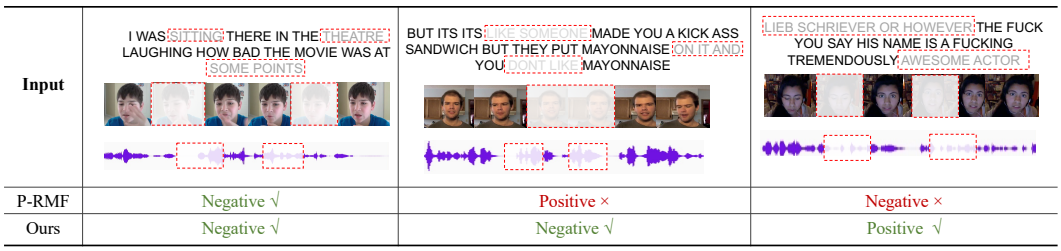

Figure 5: Case study on the MOSI testing set, where SemMSA corrects challenging examples misclassified by P-RMF under incomplete multimodal inputs.

and data protection mechanisms. Second, models trained on existing multimodal datasets may inherit demographic, cultural, linguistic, or contextual biases, potentially leading to unfair or inaccurate predictions for underrepresented groups. Third, incorrect sentiment predictions may negatively affect users if applied in high-stakes domains such as hiring and legal decision-making.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The key claims in the abstract and introduction accurately reflect the contribution and scope of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in the Limitations section of the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions and complete proofs of the theoretical results on CSA are provided in the Theoretical Analysis section of the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 and Table 9 give the implementation details and the hyperparameters of the configuration used for each metric, and these configurations are released with the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed

instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code, data, and CH-SIMS checkpoints are released at `https://github.com/peacelwh/SemMSA`, `https://huggingface.co/datasets/Bluexxx/SemMSA-data`, and `https://huggingface.co/Bluexxx/SemMSA`, with instructions for training and evaluation.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: Data splits, hyperparameters, the optimizer, and checkpoint selection are given in Section 4.1 and the Appendix (Table 9).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported; following prior works, results are obtained with random seeds 1111, 1112, and 1113.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 4.1 and the Appendix report the GPUs (NVIDIA RTX 6000 Ada and RTX 4090), and Table 6 reports trainable parameters, GFLOPs, memory, and latency.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Both potential positive and negative societal impacts are discussed in the Societal Impact section of the Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets (CMU-MOSI, CMU-MOSEI, and CH-SIMS), BERT, and Qwen3-1.7B are cited, and their licenses and terms of use are respected; the released features follow the terms of the original datasets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code (MIT License), the processed features, and the CH-SIMS checkpoints are documented in the README of the repository and on the Hugging Face dataset and model pages.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The LLM usage is deeply involved in the design of core methodology of the paper and carefully described in the manuscript and appendix.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.