Modality Disentangled Learning for Incomplete Multimodal Emotion Recognition: A Primitive Memory Distillation Perspective

Jiaqi Zhang[1], Zheng Pang[1], Mengting Li[1], Yiqi Wang[2], Guangyuan Dong[3]*, Chao Xue[4], Yusen Wu[5], Zihao Li[6], Huy Phan[7], Sicheng Zhao[8], Björn W. Schuller[9,10], Jiachen Luo[9,11]†

[1]Jiangsu University [2]Griffith University [3]National University of Singapore
[4]University of New South Wales [5]Fujian University of Technology
[6]Xi'an Jiaotong-Liverpool University [7]German Research Center for Artificial Intelligence
[8]Tsinghua University [9]Technical University of Munich
[10]Imperial College London [11]Queen Mary University of London

Abstract

Multimodal Emotion Recognition (MER) systems often suffer from missing modalities in real-world scenarios. Existing methods usually generate, align, or distill missing modalities as a whole, overlooking the heterogeneous nature of the information carried by each modality. Such holistic treatment mixes inferable shared semantics with uncertain modality-specific details, yielding unstable representations and degrading robustness. To address this issue, we propose the Primitive Memory Distillation (PriMD) framework. Unlike existing methods, PriMD takes an intra-modal perspective and focuses on how different types of information within a modality differ in recoverability. PriMD first disentangles cross-modal shared semantics from modality-specific representations, and then discretizes the latter into learnable semantic primitives to construct modality-specific memory banks. When modalities are missing, PriMD is a teacher–student framework that the student model uses the shared semantics of available modalities as queries to dynamically retrieve primitives. It compensates for missing modality-specific information within a constrained memory space and aligns with the teacher model. Extensive experiments on IEMOCAP, CMU-MOSI, and CMU-MOSEI demonstrate that PriMD achieves state-of-the-art performance and consistently stronger robustness across a wide range of missing-modality settings, while mitigating the instability caused by holistic feature inference. Our code and project website are available at `https://github.com/JiaqiZhang-Sengoku/PriMD` and `https://jiaqizhang-sengoku.github.io/PriMD/`, respectively.

1 Introduction

Multimodal Emotion Recognition (MER) aims to integrate textual, acoustic, and visual signals to understand human emotional states (Gandhi et al., 2023; Morency et al., 2011). It has broad applications in (Kirchner et al., 2019; Lu et al., 2026a,b), dialogue systems (Fu et al., 2022; Liang et al., 2022), and social media analysis (Somandepalli et al., 2021; Gu et al., 2026). Compared with unimodal inputs, multimodal inputs provide richer and complementary emotional cues. However, in real-world scenarios, privacy constraints (Aguilar et al., 2019; Zhao et al., 2021), sensor failures (Vazquez-Rodriguez et al., 2023; Wang et al., 2023b), and low-quality inputs (Wang et al., 2023b; Dong et al., 2026b) often cause some modalities to be missing, weakening the robustness of multimodal models.

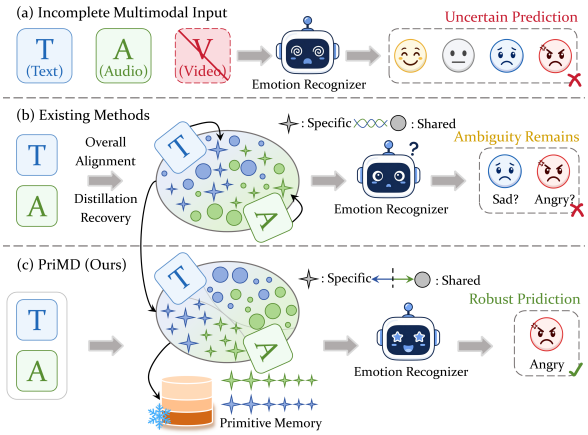


Figure 1: (a) Real-world inputs may suffer from partial modality missingness, leading to uncertain predictions. (b) Existing holistic generation, alignment, or distillation methods treat missing modalities as a whole. This tends to mix modality-shared semantics with uncertain modality-specific details, leaving the learned representations ambiguous. (c) Our PriMD first disentangles shared semantics from modality-specific information and performs constrained compensation for the missing modality-specific information. This enables more robust predictions under missing-modality conditions.

To address incomplete modalities, existing methods for incomplete multimodal emotion recognition can be broadly grouped into three categories.

*Equal Contribution with the First Author.
†Corresponding Author. jiachen.luo@qmul.ac.uk

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 47534–47552
October 24-29, 2026 ©2026 Association for Computational Linguistics

Modality completion methods recover missing features from available modalities through generators, reconstruction networks, or cross-modal mappings (Kingma and Welling, 2014; Goodfellow et al., 2014; Ho et al., 2020; Lipman et al., 2023; Wang et al., 2026a; Yang et al., 2026). However, when the missing modality contains details that are difficult to determine from other modalities, the generated features can be unreliable (Lin and Hu, 2024; Yao et al., 2024; Lin et al., 2024; Zhang et al., 2026b). Representation learning methods impose consistency constraints to map different missing-modality combinations into a unified latent space (Pham et al., 2019; Zhao et al., 2021; Lian et al., 2023; Zhang et al., 2026a). This improves model adaptability, but may suppress modality-specific information (Xu et al., 2024; Wen et al., 2024; Xu et al., 2026; Dong et al., 2026a). Knowledge distillation methods use a full-modality teacher model to guide missing-modality student models through feature, relation, or prediction-distribution distillation (Li et al., 2024a,c,b; Zhuang et al., 2025; Liu et al., 2026a,b). However, most existing distillation strategies focus on transferring teacher knowledge, without distinguishing inferable shared semantics from uncertain modality-specific details. Although these methods bring notable improvements, they still treat missing modalities as a whole during generation, alignment, or distillation. As shown in Figure 1, such holistic processing mixes inferable shared semantics with uncertain modality-specific information. This leaves ambiguity in the representations and leads to unstable predictions under missing-modality conditions.

We argue that the above issue arises because information in a missing modality is not equally recoverable. Specifically, the representation of a modality usually contains both cross-modal shared semantics and modality-specific details. The former describes emotional information that is relatively consistent across modalities and can usually be reliably estimated. The latter is tied to the expression form of a specific modality and is therefore more uncertain. Thus, when a missing modality is generated, aligned, or distilled as a whole, recoverable shared information and uncertain modality-specific details are processed together. This makes it difficult to obtain stable representations under complex missing-modality combinations. Although prior studies on full-modality MER have used feature disentanglement to mitigate modality heterogeneity (Hazarika et al., 2020; Li et al., 2023; Qian et al., 2025; Wang et al., 2026c,b; Yu et al., 2026), these methods mainly focus on representation learning with full-modality inputs and cannot recover modality-specific information under missing conditions. Based on this observation, we argue that incomplete MER should estimate shared semantics while modeling uncertain modality-specific information in a constrained manner.

To address these issues, we propose Primitive Memory Distillation (PriMD), a teacher-student framework for structured missing-modality compensation. PriMD does not treat missing-modality recovery as an isolated generation problem. Instead, it integrates recovery into structured representation learning and teacher-student knowledge transfer. In the full-modality teacher model network, we design Shared-Specific Semantic Decoupling (SSSD) to learn and disentangle cross-modal shared semantics and modality-specific representations. Discrete Primitive Memory Construction (DPMC) further discretizes modality-specific representations into learnable semantic primitives and constructs modality-specific primitive memory banks. Under missing-modality conditions, Dynamic Retrieval-Augmented Distillation (DRAD) dynamically retrieves relevant semantic primitives according to the shared semantics of available modalities. It then performs constrained compensation for missing modality-specific information and aligns the student model with the full-modality teacher, leading to stable representations and predictions. The main contributions of this paper are as follows:

- We revisit incomplete MER from an intra-modal perspective and formulate it as a unified process of shared-semantic estimation and constrained compensation of modality-specific information.
- We design SSSD to disentangle shared and modality-specific features, and DPMC/DRAD to discretize the modality-specific space into semantic primitives and retrieve them for compensation under missing modalities.
- Extensive experiments on three benchmarks show that PriMD outperforms prior methods under diverse missing-modality settings, with especially large gains when two modalities are absent.

2 Related Work

Multimodal Learning in MER. Multimodal Emotion Recognition (MER) aims to learn discriminative emotion representations from heterogeneous signals, such as text, speech, and vision. Early methods mainly focused on modality fusion. TFN (Zadeh et al., 2017) models unimodal and cross-modal interactions through tensor outer products, while LMF (Liu et al., 2018) uses low-rank factorization to reduce computational cost. As sequential context and modality asynchrony received increasing attention, later methods further introduced memory mechanisms and cross-modal attention. MFN (Zadeh et al., 2018a) introduces multi-view memory units to model cross-modal interactions over time, whereas MulT (Tsai et al., 2019) adopts directional cross-modal attention to handle unaligned multimodal sequences. To alleviate modality redundancy, noise interference, and semantic inconsistency, recent studies have started to improve model robustness at the representation level. MISR (Zhu et al., 2025a) mitigates bias in emotion distributions through invariance learning, while DecAlign (Qian et al., 2025) models modality-shared semantics and modality-specific components, preserving both semantic consistency and modality differences. While effective, these methods presuppose full-modality inputs and cannot be directly applied when modalities are absent.

Incomplete Multimodal Learning in MER. Missing modalities are a common issue in real-world multimodal emotion recognition. They are usually caused by sensor failures, noise interference, privacy constraints, or automatic recognition errors. Existing methods can be broadly categorized into three categories: modality completion, robust representation learning, and knowledge distillation. CRA (Tran et al., 2017) models modality correlations through cascaded residual autoencoders, while IMDer (Wang et al., 2023d) uses a score-based diffusion model to generate missing features. MPLMM (Guo et al., 2024) and P-RMF (Zhu et al., 2025b) enhance missing-information reconstruction through prompt learning and latent proxy modalities, respectively. Representation learning methods avoid explicit generation and instead learn the joint distribution of modalities. MCTN (Pham et al., 2019) learns cross-modal representations through cyclic translation, MMIN (Zhao et al., 2021) uses cycle-consistency constraints to compensate for missing information, and GCNet (Lian et al., 2023) further introduces graph neural networks to model temporal dependencies. In recent years, knowledge distillation has also been adopted to transfer supervision from full-modality teachers models to missing-modality students models (Gou et al., 2021; Yang et al., 2025). For example, UMDF (Li et al., 2024a), CorrKD (Li et al., 2024c), and HRLF (Li et al., 2024b) improve knowledge transfer from the perspectives of self-distillation, correlation distillation, and hierarchical mutual-information alignment, respectively. However, most existing methods focus on the holistic generation, alignment, and distillation of missing modalities, while paying limited attention to the characterization of modality-specific and shared information. When available modalities cannot reliably infer the missing specific details, such holistic processing may introduce irrelevant or even conflicting features.

3 Methodology

Problem Formulation. This paper considers a multimodal emotion recognition task with three modalities: textual, acoustic, and visual signals. Let the full-modality set be $\Omega = \{t, a, v\}$. For a given sample, its multimodal input is represented as $X = \{X^m \mid m \in \Omega\}$, where $X^m = \{x_i^m\}_{i=1}^{T_m}$ denotes the sequential features of modality m. In the incomplete-modality setting, only a subset of modalities is available to the model. We use $O \subseteq \Omega$ to denote the available modality set and $M = \Omega \backslash O$ to denote the missing modality set, where $O \neq \emptyset$, $O \cup M = \Omega$, and $O \cap M = \emptyset$. Accordingly, $X_O = \{X^m \mid m \in O\}$ denotes the available modality input, while $X_M = \{X^m \mid m \in M\}$ denotes the missing-modality feature set. To represent different missing-modality combinations, we use a modality mask $c = [c_t, c_a, c_v]$, where $c_m = 1$ indicates that modality m is available, and $c_m = 0$ indicates that it is missing.

Overview. As shown in Figure 2, the proposed PriMD consists of a full-modality teacher model and an incomplete-modality student model. The teacher model first disentangles complete multimodal representations into cross-modal shared semantics and modality-specific information, and then constructs a discrete primitive memory based on the modality-specific information. In the student model, when some modalities are missing, the model dynamically retrieves relevant primitives according to the shared semantics of the available

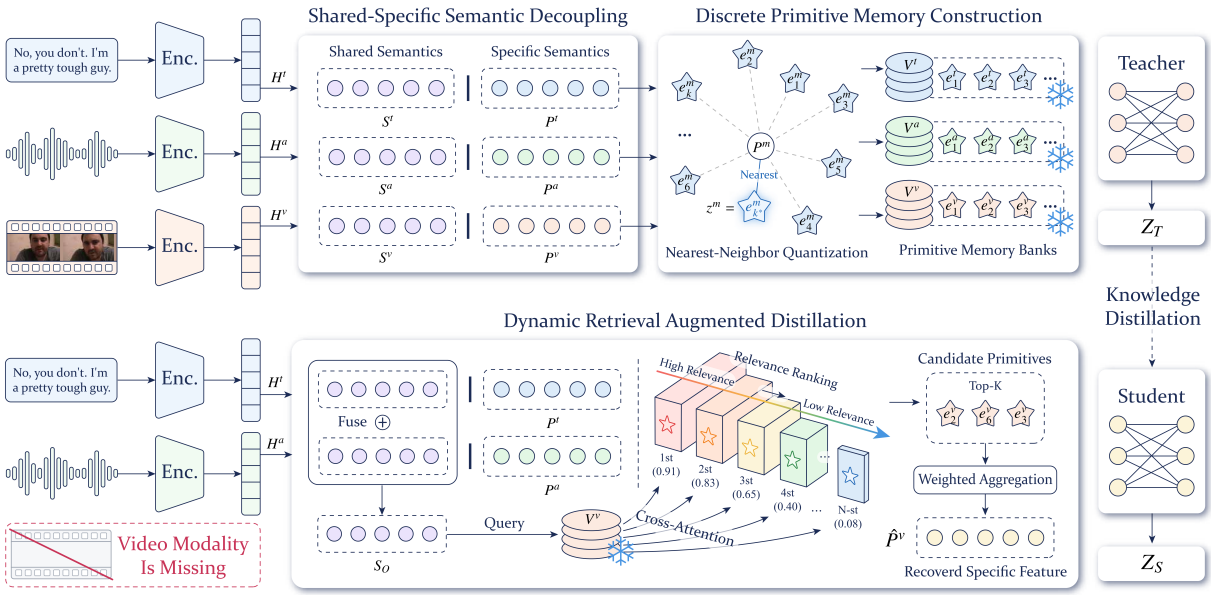


Figure 2: **Overall framework of PriMD.** The full-modality teacher model first encodes the inputs and disentangles shared semantics from modality-specific information via SSSD. Then, DPMC quantizes the modality-specific representations into discrete semantic primitives and constructs primitive memory banks. In missing-modality scenarios, the student model uses DRAD to dynamically retrieve relevant primitives from the memory banks based on the semantics of available modalities, enabling constrained compensation for the missing information. Finally, the compensated student model representations are aligned with the teacher model through retrieval-augmented distillation.

modalities to compensate for the missing modality-specific information. Finally, PriMD aligns the student model with the full-modality teacher model via retrieval-augmented distillation.

3.1 Shared-Specific Semantic Decoupling

Multimodal representations usually contain both cross-modal shared semantics and modality-specific information. Effectively disentangling these two types of information can reduce redundant feature interference in subsequent knowledge transfer and provide a more stable representation basis for missing-modality compensation. To this end, we first use the full-modality teacher network F_T to encode the input of each modality. Given the complete input X^m, the modality encoder extracts deep features:

$$H^m = \psi^m_{\text{enc}}(X^m) \in \mathbb{R}^{N\times d}, \tag{1}$$

where N denotes the number of samples in a mini-batch. Then, two independent projection branches map H^m into the shared semantic representation S^m and the modality-specific representation P^m:

$$S^m = \psi_s(H^m) \in \mathbb{R}^{N\times d}, \tag{2}$$

$$P^m = \psi_p(H^m) \in \mathbb{R}^{N\times d}. \tag{3}$$

To encourage S^m to capture cross-modal shared information, we minimize the distance between shared representations from different modalities:

$$L_{\text{Align}} = \frac{2}{N|\Omega|(|\Omega|-1)} \sum_{m\in\Omega} \sum_{n>m} \|S^m - S^n\|_F^2 . \tag{4}$$

Meanwhile, to suppress the statistical dependence between the shared semantic representation and the modality-specific representation, and to make P^m focus more on modality-specific information, we map the features into a reproducing kernel Hilbert space and use the Hilbert-Schmidt Independence Criterion as a constraint:

$$L_{\text{Indep}} = \sum_{m\in\Omega} \frac{1}{(N-1)^2} \operatorname{Tr}\left(K_S^m C_N K_P^m C_N\right), \tag{5}$$

where $C_N = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ denotes the centering matrix, and $\operatorname{Tr}(\cdot)$ denotes the matrix trace. K_S^m and K_P^m are the Gram matrices of S^m and P^m under the radial basis function kernel mapping, respectively. $I_N \in \mathbb{R}^{N\times N}$ is the identity matrix, and $\mathbf{1}_N \in \mathbb{R}^N$ is an all-one vector. After disentanglement, the teacher network concatenates the shared and modality-specific features of all modalities and outputs Z_T through the classifier Φ^T_{Cls}:

$$Z_T = \Phi^T_{\text{Cls}}(\text{Concat}_{m\in\Omega}\,(S^m, P^m)). \tag{6}$$

3.2 Discrete Primitive Memory Construction

Continuous high-dimensional feature spaces usually exhibit large distributional variance. When modalities are missing, directly regressing such continuous features is often difficult to optimize and may introduce strong uncertainty. To reduce the difficulty of missing-information compensation and provide more reliable prior representations, we discretize the disentangled modality-specific representations into a finite number of semantic primitives and construct modality-specific primitive memory banks.

For modality m, we define its learnable primitive memory bank as:

$$V^m = \{e_1^m, e_2^m, \ldots, e_C^m\} \in \mathbb{R}^{C\times d}, \tag{7}$$

where C denotes the primitive capacity. Given the modality-specific representation $P^m = \{p_i^m\}_{i=1}^N \in \mathbb{R}^{N\times d}$, we quantize each sample representation p_i^m to the nearest semantic primitive in the memory bank:

$$k_i^* = \arg\min_{1\le k\le C} \|p_i^m - e_k^m\|_2^2, \quad z_i^m = e_{k_i^*}^m, \tag{8}$$

where all quantized primitive representations form $Z^m = \{z_i^m\}_{i=1}^N \in \mathbb{R}^{N\times d}$. To make the discrete primitives in the memory bank fit the distribution of modality-specific representations while preventing continuous representations from deviating excessively from their matched primitives during training, we use a vector quantization loss with a commitment constraint:

$$\begin{aligned} L_{\mathrm{VQ}} = \sum_{m\in\Omega} \Big(&\|\mathrm{sg}(P^m) - Z^m\|_F^2 \\ &+ \beta \|P^m - \mathrm{sg}(Z^m)\|_F^2 \Big), \end{aligned} \tag{9}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation, and β is a tunable parameter. After the teacher stage is completed, the modality-specific primitive memory banks are frozen and used as prior retrieval spaces in the student stage.

3.3 Dynamic Retrieval-Augmented Distillation

When modalities are missing, the student network lacks complete input cues. Directly aligning it with the full-modality teacher model may introduce incorrect or unstable feature matching. To address this issue, we retrieve from the frozen primitive memory banks to provide constrained feature compensation for missing modalities. Moreover, considering differences in sample difficulty and missing-modality combinations, a fixed number of retrieved primitives is insufficient for complex scenarios (Song et al., 2023). Therefore, we design a dynamic retrieval-augmented distillation strategy.

For each missing modality $m \in M$, the student network F_S first extracts and fuses shared semantics from the available modality set O:

$$S_O = \mathrm{Fuse}\left(\{S^o \mid o \in O\}\right) \in \mathbb{R}^{N\times d}. \tag{10}$$

where $\mathrm{Fuse}(\cdot)$ denotes the available-modality shared-semantics fusion operator. Then, using S_O as the query, we compute the relevance scores between S_O and the primitives in the memory bank V^m of the missing modality m:

$$A^m = \mathrm{Softmax}\left(\frac{(S_O W_q)(V^m W_k)^\top}{\sqrt{d_r}}\right) \in \mathbb{R}^{N\times C}. \tag{11}$$

Here, W_q and W_k map the shared semantics and primitives into the same retrieval space, respectively. d_r is the dimension of the retrieval space, and $\mathrm{Softmax}(\cdot)$ is applied along the primitive dimension. For the b-th sample, we sort the primitives in the memory bank in descending order according to its relevance scores A_b^m. The primitive ranked at position k is denoted as $\tilde{e}_{b,k}^m$, and its attention weight is denoted as $\tilde{a}_{b,k}^m$. To dynamically determine the number of retrieved primitives, we use $\alpha^m = \mathrm{MLP}(S_O) \in \mathbb{R}^{N\times K_{\max}}$ to predict retrieval-number logits for the missing modality m, and then generate a differentiable retrieval-number distribution:

$$\pi_{b,r}^m = \frac{\exp\Big((\alpha_{b,r}^m + g_{b,r}^m)/\gamma\Big)}{\sum_{j=1}^{K_{\max}} \exp\Big((\alpha_{b,j}^m + g_{b,j}^m)/\gamma\Big)}, \tag{12}$$

where $g_{b,r}^m \sim \mathrm{Gumbel}(0,1)$, γ is a hyperparameter controlling the smoothness of the distribution, and $r \in \{1,\ldots,K_{\max}\}$ denotes a candidate retrieval number. For candidate retrieval number r, we aggregate the top-r primitives and normalize their attention weights:

$$\phi_{b,r}^m = \frac{\sum_{k=1}^r \tilde{a}_{b,k}^m \tilde{e}_{b,k}^m}{\sum_{k=1}^r \tilde{a}_{b,k}^m + \epsilon}. \tag{13}$$

Finally, the compensated representation of the missing modality m for the b-th sample is obtained as:

$$\hat{P}^m = \left(\sum_{r=1}^{K_{\max}} \pi_{b,r}^m \phi_{b,r}^m\right)_{b=1}^N \in \mathbb{R}^{N\times d}. \tag{14}$$

After obtaining the compensated representations, the student network concatenates the shared semantics of the available modalities, the modality-specific representations of the available modalities, and compensated missing-modality representations. The concatenated representation is fed into the classifier Φ^S_{Cls} to produce the student logits:

$$Z_S = \Phi^S_{\text{Cls}} \Big(\text{Concat} \Big(S_O, \underset{o \in O}{\text{Concat}}\, P^o, \underset{m \in M}{\text{Concat}}\, \hat{P}^m \Big) \Big). \tag{15}$$

To prevent the model from always selecting the maximum retrieval number, we introduce a nonlinear cost penalty for retrieval:

$$L_{\text{Cost}} = \frac{1}{N} \sum_{b=1}^{N} \sum_{m \in M} \sum_{r=1}^{K_{\max}} \pi^m_{b,r} (\exp(\frac{r}{K_{\max}}) - 1). \tag{16}$$

Finally, we use the output Z_T of the frozen full-modality teacher model as soft labels to distill knowledge into the student model:

$$L_{\text{KD}} = \tau^2_{\text{KD}} \cdot D_{\text{KL}} \Big(\text{Softmax}(Z_T / \tau_{\text{KD}}) \Big\| \text{Softmax}(Z_S / \tau_{\text{KD}}) \Big), \tag{17}$$

where τ_{KD} is the distillation temperature.

3.4 Joint Optimization

We train PriMD with a two-stage optimization strategy. In the teacher stage, we use full-modality data to optimize the full-modality teacher network F_T and the modality-specific primitive memory banks V. After training, the teacher network and memory banks are frozen and used as the knowledge source and retrieval space for the student stage. In the student stage, we train the student network F_S with missing-modality masks, enabling it to perform dynamic retrieval, feature compensation, and knowledge alignment under incomplete-modality conditions. The optimization objectives of the two stages are defined as:

$$L_{\text{Tea}} = L^T_{\text{Task}} + \lambda_{\text{Align}} L_{\text{Align}} + \lambda_{\text{Indep}} L_{\text{Indep}} + \lambda_{\text{VQ}} L_{\text{VQ}}, \tag{18}$$

$$L_{\text{Stu}} = L^S_{\text{Task}} + L_{\text{KD}} + L_{\text{Cost}}. \tag{19}$$

where λ_{Align}, λ_{Indep}, and λ_{VQ} weight the corresponding losses.

4 Experiments

4.1 Experiment Setup

Datasets. To evaluate the effectiveness of our proposed method, we conduct extensive experiments on three datasets: IEMOCAP (Busso et al., 2008), CMU-MOSI (Zadeh et al., 2016), and CMU-MOSEI (Zadeh et al., 2018b).

Evaluation Metrics. For IEMOCAP, we report Weighted Accuracy (WA) and Unweighted Accuracy (UA). For CMU-MOSI and CMU-MOSEI, we adopt task-specific metrics, including Accuracy (ACC), F1-score, and Mean Absolute Error (MAE), following the corresponding experimental settings.

Details of the datasets and evaluation metrics are shown in Appendix A.

Implementation Details. Detailed experimental setup and parameters are provided in Appendix B.

4.2 Comparison with SOTA Methods

Comparison under inter-modal missingness. To evaluate the robustness of PriMD in incomplete multimodal scenarios, we quantitatively compare it with SOTA methods under different missing-modality settings, as shown in Table 1. PriMD achieves the best average performance across all three experimental settings in terms of the average over the six missing-modality combinations. Specifically, on the four-class IEMOCAP task, PriMD achieves 77.43% WA and 77.28% UA, outperforming the second-best method, HARDY-MER, by 1.87% and 1.86%, respectively. On the more challenging six-class IEMOCAP task, PriMD achieves 58.13% WA and 54.62% UA, improving over HARDY-MER by 1.42% and 1.97%, respectively. On CMU-MOSEI, PriMD obtains an average ACC/F1 of 79.81%/79.57%, also surpassing existing methods. Overall, these results indicate that holistic reconstruction or direct alignment of missing modalities is susceptible to uncertain modality-specific information. In contrast, PriMD performs constrained compensation for missing-specific information through shared-semantics-guided primitive retrieval, thereby more effectively mitigating representation shifts caused by missing modalities.

Comparison for Intra-modal Missingness: We further evaluate the robustness of PriMD under intra-modal missingness on CMU-MOSI and CMU-MOSEI. Following previous works (Zhang et al., 2024; Zhu et al., 2025b), we gradually increase the missing rate from 0.0 to 0.9. As shown in Figure 3, the performance of most methods de-

Table 1: Performance comparison with SOTA methods on two datasets under various missing-modality scenarios. "Avg." indicates the mean performance across all settings. The best and suboptimal results are highlighted. A, T, and V denote the acoustic, textual, and visual modalities, respectively. Gray shading indicates the presence of a modality, while gray dashed lines indicate its absence.

Dataset	Method	A	T	V										Avg.	
		WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)
IEMOCAP Four Class	CPMNet (Zhang et al., 2022)	46.85	51.72	45.63	45.32	44.95	44.49	34.81	36.23	48.67	49.33	45.62	46.57	44.42	45.61
	MMIN (Zhao et al., 2021)	56.58	59.00	66.57	68.02	52.52	50.60	72.94	71.14	63.99	63.43	71.67	68.61	64.05	63.47
	GCNet (Lian et al., 2023)	65.58	68.76	72.33	70.42	57.96	52.54	77.02	76.87	67.40	65.64	75.63	73.62	69.32	67.98
	CIF-MMIN (Liu et al., 2024)	57.53	60.06	67.22	68.99	53.46	51.56	74.19	72.59	64.99	63.53	72.40	69.91	64.97	64.44
	MoMKE (Xu et al., 2024)	69.53	70.21	77.30	77.66	56.80	52.03	79.03	79.88	68.57	66.22	75.55	74.18	71.13	70.03
	HARDY-MER (Liu et al., 2025)	72.65	73.87	82.49	82.69	63.19	60.54	81.67	82.43	74.19	74.50	79.18	78.51	75.56	75.42
	PriMD (Ours)	73.82	74.46	84.13	84.09	64.77	62.81	83.65	84.83	75.06	74.84	83.17	82.66	77.43	77.28
IEMOCAP Six Class	CPMNet (Zhang et al., 2022)	29.47	29.80	32.44	34.95	26.20	24.95	33.49	33.94	26.92	25.46	31.34	30.43	29.98	29.92
	MMIN (Zhao et al., 2021)	44.08	42.96	42.17	38.55	35.74	30.65	51.95	48.31	41.92	38.15	47.49	40.63	43.89	39.88
	GCNet (Lian et al., 2023)	49.95	46.45	56.48	55.62	39.78	34.97	58.24	57.25	47.57	43.31	57.43	54.66	51.58	48.71
	CIF-MMIN (Liu et al., 2024)	44.96	43.56	43.40	39.71	36.11	31.35	52.43	49.20	42.54	39.22	48.88	44.91	44.72	41.33
	MoMKE (Xu et al., 2024)	50.51	47.38	61.09	60.19	39.07	34.51	63.18	61.94	48.65	44.08	59.92	57.55	53.74	50.94
	HARDY-MER (Liu et al., 2025)	51.58	49.14	65.89	61.95	43.02	36.49	65.18	62.98	52.91	47.45	61.66	57.86	56.71	52.65
	PriMD (Ours)	53.62	51.43	66.21	62.17	45.35	39.86	65.78	63.66	54.17	49.82	63.65	60.77	58.13	54.62
		ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)
CMU MOSEI	CPMNet (Zhang et al., 2022)	65.71	65.18	72.87	72.44	61.23	61.73	72.65	72.24	61.56	61.99	66.29	66.84	66.72	66.74
	MMIN (Zhao et al., 2021)	58.90	59.50	82.20	82.40	59.30	60.01	83.70	83.30	63.55	61.91	81.75	81.42	71.57	71.42
	GCNet (Lian et al., 2023)	72.04	70.34	84.26	84.17	68.08	67.25	85.10	85.10	71.49	69.96	84.74	84.54	77.62	76.89
	CIF-MMIN (Liu et al., 2024)	63.87	64.60	83.53	83.04	61.96	62.66	84.01	83.47	64.68	62.08	82.50	81.94	73.43	72.97
	MoMKE (Xu et al., 2024)	72.56	71.03	86.10	86.03	64.50	63.46	86.32	86.29	72.37	72.07	86.90	86.91	78.13	77.63
	CMAD (Zhuang et al., 2025)	63.00	60.80	86.10	86.00	65.70	64.40	86.30	86.20	65.60	64.80	86.40	86.40	75.52	74.77
	HARDY-MER (Liu et al., 2025)	74.82	74.11	87.20	87.13	69.35	67.50	85.42	85.01	74.82	74.11	85.72	85.39	79.56	78.88
	PriMD (Ours)	74.61	74.25	87.36	87.24	68.96	67.36	87.26	87.69	73.85	73.61	86.81	87.27	79.81	79.57

creases as the missing rate increases, while PriMD maintains higher F1 scores and lower MAE under most missing-rate settings. In contrast, P-RMF mitigates the information deficiency via proxy modalities, but its compensation still operates on holistic representations and therefore making it vulnerable to uncertain details when intra-modal information is continuously lost. For LNLN, its F1 score improves under some settings as the missing rate increases, whereas its MAE continues to increase. This suggests that its classification results may be affected by data bias, making the model more likely to predict high-frequency classes rather than learn stable affective representations. PriMD shows more stable performance across different missing rates, indicating that it can capture shared semantics from the remaining modalities and reduce the representation shift caused by intra-modal information loss.

4.3 Ablation Study

To verify the effectiveness of the core modules in PriMD, we conducted ablation experiments. The results are shown in Table 2.

1) w/o SSSD: To evaluate the role of shared-specific semantic disentanglement, we replace the two projection branches with a single projection, so that each modality yields one entangled representation, without separating cross-modal shared semantics from modality-specific information. The average performance of this variant drops to 74.36% WA and 74.05% UA. The results show that directly mixing these two types of information for subsequent memory construction and distillation makes the model vulnerable to redundant or unstable modality-specific information, thereby weakening its representation robustness under missing-modality scenarios.

2) w/o DPMC: To analyze the contribution of discrete primitive memory construction, we removed DPMC. Compared with the full model, this variant decreases the average WA/UA to 75.95%/75.76%, respectively. This indicates that modality-specific information in the continuous feature space is uncertain, and directly performing generation or alignment makes it difficult to obtain stable representations. DPMC provides a more constrained compensation space for missing specific information through primitive memory.

3) w/o DRAD: To verify the contribution of dynamic retrieval-augmented distillation, we remove the entire DRAD module. In this variant, the student model uses only the shared and modality-specific representations of the available modalities. The results show that this setting leads to the most severe performance degradation, with average results of only 72.31% WA and 72.10% UA. This indicates that representations from the available

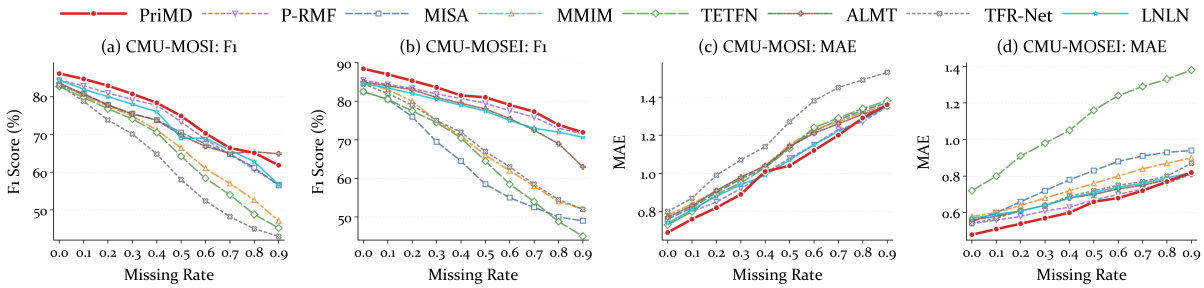


Figure 3: Performance curves of different methods under varying missing rates on the CMU-MOSI and CMU-MOSEI datasets. Subfigures (a)-(b) and (c)-(d) show F1 and MAE results on MOSI and MOSEI, respectively.

Table 2: Results of the ablation experiments under the Four-Class IEMOCAP. The best results are highlighted.

Dataset	Modules	A T V		A T V		A T V		A T V		A T V		A T V		Avg.	
		WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)
IEMOCAP Four Class	w/o SSSD	70.15	70.82	81.25	81.02	61.34	59.45	80.77	81.12	72.41	72.05	80.22	79.85	74.36	74.05
	w/o DPMC	72.41	73.15	82.88	82.91	63.12	61.05	81.95	82.74	73.88	73.61	81.45	81.12	75.95	75.76
	w/o DRAD	69.88	70.02	79.45	79.12	58.76	57.11	78.33	79.65	69.52	69.18	77.94	77.53	72.31	72.10
	PriMD (Ours)	73.82	74.46	84.13	84.09	64.77	62.81	83.65	84.83	75.06	74.84	83.17	82.66	77.43	77.28

modalities alone are insufficient to adapt to different missing-modality combinations.

4.3.1 Decomposition and Compensation

As shown in Table 3(a), we constructed a capacity-matched variant, Holistic-PriMD, to evaluate the role of shared-specific semantic disentanglement. This variant constructs the memory banks directly from holistic modality representations and queries them using the fused representation of the available modalities. Compared with Holistic-PriMD, PriMD improved WA/UA from 75.02%/74.81% to 77.43%/77.28%, while reducing cross-combination JS and teacher–student KL from 0.091/0.121 to 0.066/0.081. These results indicate that shared-specific semantic disentanglement improves recognition performance and reduces prediction shifts across missing-modality combinations and teacher–student discrepancies under matched model capacity.

As shown in Table 3(b), we further separated the contribution of semantic disentanglement from that of the compensation mechanism by comparing SSSD+MLP, SSSD+CVAE, continuous compensation, and discrete primitive retrieval under the same SSSD framework. PriMD achieved the highest WA/UA of 77.43%/77.28% and the lowest JS/KL of 0.066/0.081. Moreover, its normalized within-class dispersion was 0.64, compared with 1.00 for continuous compensation. These results indicate that discrete primitive retrieval provides an independent benefit under the same disentanglement framework and is associated with a more compact within-class distribution.

4.3.2 SSSD Components

4) Ablation Study on SSSD: To further verify the role of different constraints in SSSD, we separately remove the shared semantic alignment loss L_{Align} and the independence constraint L_{Indep}. As shown in Table 5, using only L_{Align} or L_{Indep} leads to lower performance than the full setting. This indicates that L_{Align} helps improve the consistency of shared semantics across modalities, while L_{Indep} reduces redundant correlations between shared and modality-specific representations. By combining them, the model achieves more effective disentanglement.

4.4 Computational Efficiency

As shown in Table 4, PriMD reduces the parameter count, training time, and peak CUDA memory by 79.7%, 51.4%, and 82.2%, respectively, compared with HARDY-MER, while increasing inference latency by 8.1%.

4.5 Hyperparameter Sensitivity

We evaluate the sensitivity of PriMD to the primitive capacity C and β in DPMC. As shown in Figure 4, PriMD remains stable across a wide range of settings. A smaller C limits the diversity of specific primitives, while an excessively large C may introduce redundant primitives and reduce retrieval reliability. Similarly, a smaller β provides

Table 3: Analysis of shared-specific decomposition and feature compensation on Four-Class IEMOCAP. Disp. denotes normalized within-class dispersion. The best results are highlighted.

(a) Shared-specific decomposition	WA (%)	UA (%)	Cross-combination JS ↓	Teacher–student KL ↓
Holistic-PriMD	75.02 ± 0.58	74.81 ± 0.61	0.091 ± 0.004	0.121 ± 0.006
PriMD (Ours)	77.43 ± 0.43	77.28 ± 0.47	0.066 ± 0.003	0.081 ± 0.004

(b) Feature compensation	WA (%)	UA (%)	Disp. ↓	Cross-combination JS ↓	Teacher–student KL ↓
SSSD+MLP	75.98 ± 0.52	75.81 ± 0.56	–	0.089 ± 0.005	0.118 ± 0.007
SSSD+CVAE	76.27 ± 0.49	76.08 ± 0.53	–	0.082 ± 0.004	0.109 ± 0.006
Continuous compensation	76.34 ± 0.48	75.83 ± 0.52	1.00	0.079 ± 0.004	0.106 ± 0.006
PriMD (Ours)	77.43 ± 0.43	77.28 ± 0.47	0.64	0.066 ± 0.003	0.081 ± 0.004

Table 4: Computational efficiency comparison.

Method	Params.	Train (s)	Memory (MiB)	Infer. (ms)
HARDY-MER	11.52M	90.981	1839.4	26.829
PriMD (Ours)	2.34M	44.205	327.4	29.013

Table 5: Ablation results of SSSD components.

SSSD		IEMOCAP Four Class		CMU MOSEI	
L_{Align}	L_{Indep}	WA(%)	UA(%)	ACC(%)	F1(%)
✓		75.39	75.87	77.82	78.06
	✓	75.96	76.24	77.58	77.64
✓	✓	77.43	77.28	79.81	79.57

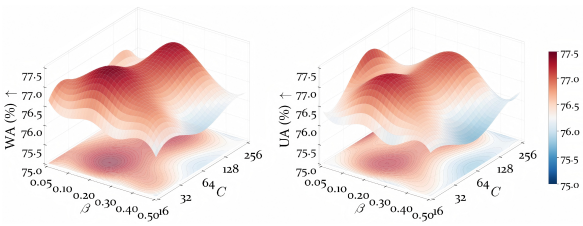


Figure 4: Parameter sensitivity on C and β of PriMD.

insufficient constraints between modality-specific features and discrete primitives, whereas a larger β over-constrains the representation space. These results show that PriMD benefits from a primitive space with balanced constraints for missing-modality compensation.

4.6 Maximum Retrieval Size

We further analyze the effect of $K_{\max}$, which controls the maximum number of primitives retrieved by DRAD. As shown in Figure 5, a smaller $K_{\max}$ restricts the student model to very limited modality-specific evidence. Increasing $K_{\max}$ allows the model to consider more complementary primitives and improves the robustness of compensation. However, an excessively large $K_{\max}$ may introduce weakly relevant primitives and dilute the reconstructed representation.

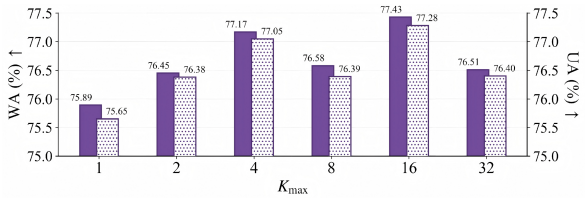


Figure 5: Performance under different K_{max}.

5 Conclusion

In this paper, we proposed PriMD, a primitive memory distillation framework for incomplete multimodal emotion recognition. Unlike holistic modality-inference methods, PriMD first decomposes multimodal representations into cross-modal shared semantics and modality-specific information, and then discretizes the latter into modality-specific primitive memories. Under missing-modality conditions, the student model leverages the shared semantics of available modalities to retrieve relevant primitives. This enables constrained compensation for uncertain modality-specific cues and effective distillation from the full-modality teacher model. Extensive experiments showed that PriMD achieves robust performance under both inter-modal and intra-modal missingness. It effectively alleviates the instability caused by holistic feature inference and provides a new perspective for robust incomplete multimodal learning.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback.

Limitations

Although PriMD achieves good robustness under various incomplete multimodal settings, several limitations remain. First, PriMD relies on a full-modality teacher model to learn shared-specific representations and construct modality-specific primitive memory banks. When full-modality training samples are scarce, biased, or costly to collect, the learned primitives may not fully cover the modality-specific information in the target domain. To reduce this dependence, future work may combine cross-modal self-supervised pretraining or semi-supervised learning to make better use of incomplete samples. Confidence-weighted primitive updates may also allow high-confidence incomplete samples to expand the memory banks. Domain adaptation could further reduce the distribution gap between the training and target domains. Second, our current experiments mainly use sample-level fixed missing masks to simulate inter-modal and intra-modal missingness. In real applications, the set of available modalities may change over time. For example, audio or video sensors may fail temporarily, and input quality may gradually decrease during continuous interaction. Such online missingness can be represented by a time-varying modality mask. PriMD does not yet explicitly model the dependence between missing states at different time steps. Future work will explore time-aware retrieval, modality-quality gating, historical primitive caching, and online memory updates. These methods may allow the model to continuously adjust its compensation strategy according to real-time modality availability and quality, improving reliability and generalization in practical deployment.

References

Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang. 2019. Multimodal and multi-view models for emotion recognition. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 991–1002. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Guangyuan Dong, Chuang Liu, Yangchen Zeng, Haoyu Wang, Xiaoyang Yu, Pinlong Zhao, Yuchao Hou, Ziwei Li, and Zheng Lin. 2026a. When generated images look right and retrieve wrong: Coverage-guided cross-scale re-indexing for knowledge-faithful generative perception. *arXiv preprint arXiv:2608.20810*.

Guangyuan Dong, Lingdong Shen, Xudong Zhang, Zonghao Ying, Haodong Jing, Ziyu Song, Shenghao Liu, Jiachen Luo, Hao Zhang, Haoyu Wang, and Zihao Li. 2026b. Lance: A LANGUAGE-ANCHORED CLOSED-LOOP ESTIMATOR for bridging low-level image fusion and high-level semantic understanding in remote sensing. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management*, CIKM '26.

Tingchen Fu, Sheng Gao, Xueliang Zhao, Ji-Rong Wen, and Rui Yan. 2022. Learning towards conversational AI: A survey. *AI Open*, 3:14–28.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26866–26875. IEEE.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *CoRR*, abs/1406.2661.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.

Bingbing Gu, Saihua Cai, Jing Wang, Zhuole Li, Xiheng Jia, and Jiaqi Zhang. 2026. Dg-scl: Diffusion-guided semantic contrastive learning for imbalanced malicious traffic detection. *Information Sciences*, page 123851.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1726–1736. Association for Computational Linguistics.

Wei Han, Hui Chen, and Soujanya Poria. 2021. 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*,

pages 9180–9192. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *ACM International Conference on Multimedia*, pages 1122–1131. ACM.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations*.

Elsa Andrea Kirchner, Stephen H. Fairclough, and Frank Kirchner. 2019. Embedded multimodal interfaces in robotics: applications, future trends, and societal implications. In *The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions*. Association for Computing Machinery.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. 2024a. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *AAAI Conference on Artificial Intelligence*, pages 10074–10082. AAAI Press.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. 2024b. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *Advances in Neural Information Processing Systems*.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024c. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468. IEEE.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640. IEEE.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2022. Emotional conversation generation with heterogeneous graph neural network. *Artificial Intelligence*, 308:103714.

Ronghao Lin and Haifeng Hu. 2024. Adapt and explore: Multimodal mixup for representation learning. *Information Fusion*, 105:102216.

Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Wenzhong Tang, Zitong Yu, and Alex C. Kot. 2024. Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 211–221. IEEE.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *International Conference on Learning Representations*. OpenReview.net.

Bingchen Liu, Yuanyuan Fang, Lei Liu, Guangyuan Dong, Xing Fu, Yuanyuan Gao, Shuyue Wei, Xin Li, and Xiangtian Meng. 2026a. Conmem: Contribution-aware memory for long-horizon manufacturing inspection logs. *arXiv preprint arXiv:2607.28126*.

Jiaxin Liu, Ding Zhong, Yue Wang, Zhidong Yang, Zhaolu Kang, Guangyuan Dong, Qishi Zhan, Pengcheng Fang, and Aofan Liu. 2026b. Dual-pathway circuits of object hallucination in vision-language models. *arXiv preprint arXiv:2605.13156*.

Rui Liu, Haolin Zuo, Zheng Lian, Björn W. Schuller, and Haizhou Li. 2024. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873.

Rui Liu, Haolin Zuo, Zheng Lian, Hongyu Yuan, and Qi Fan. 2025. Hardness-aware dynamic curriculum learning for robust multimodal emotion recognition with missing modalities. In *ACM International Conference on Multimedia*, pages 5755–5764. ACM.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2247–2256. Association for Computational Linguistics.

Shuo Lu, Jianjie Cheng, Yinuo Xu, Yongcan Yu, Lijun Sheng, Peijie Wang, Siru Jiang, Yongguan Hu, Run Ling, Yihua Shao, et al. 2026a. Do mllms really understand space? a mathematical reasoning evaluation. *arXiv preprint arXiv:2602.11635*.

Shuo Lu, Yinuo Xu, Jianjie Cheng, Lingxiao He, Meng Wang, and Jian Liang. 2025. Deepresearch-slice: Bridging the retrieval-utilization gap via explicit text slicing. *arXiv preprint arXiv:2601.03261*.

Shuo Lu, Yinuo Xu, Kecheng Yu, Siru Jiang, Yongcan Yu, Yubin Wang, Haitao Yang, Yuxiang Zhang, Bin Wang, Ran He, et al. 2026b. Worldcoder-bench: Benchmarking physically grounded 3d world synthesis. *arXiv preprint arXiv:2606.01869*.

Wei Luo, Mengying Xu, and Hanjiang Lai. 2023. Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In *MultiMedia Modeling*, Lecture Notes in Computer Science, pages 411–422. Springer.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *AAAI Conference on Artificial Intelligence*, pages 164–172. AAAI Press.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive rnn for emotion detection in conversations. In *AAAI Conference on Artificial Intelligence*, pages 6818–6825. AAAI Press.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *ICMI*, pages 169–176. ACM.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI Conference on Artificial Intelligence*, pages 6892–6899. AAAI Press.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. 2025. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. *CoRR*, abs/2503.11892.

Pin Qian, Su Wang, Yihang Chen, Qiaolin Yu, Xiaoyuan Wang, Zhitong Guo, Zhicheng Wang, and Junxian You. 2026a. Toward user-conditioned evaluation of personal llm agents under temporal interventions. *arXiv preprint arXiv:2607.21635*.

Pin Qian, Su Wang, Xiaoyuan Wang, Yihang Chen, Wenxuan Xu, Qiaolin Yu, Shuhuai Lin, Sipeng Zhang, Junxian You, and Xinpeng Wei. 2026b. Relevant is not warranted: Evidence-force calibration for cited rag. *arXiv preprint arXiv:2605.28044*.

Krishna Somandepalli, Tanaya Guha, Victor R. Martinez, Naveen Kumar, Hartwig Adam, and Shrikanth Narayanan. 2021. Computational media intelligence: Human-centered machine analysis of media. *Proceedings of the IEEE*, 109(5):891–910.

Zitao Song, Yining Wang, Pin Qian, Sifan Song, Frans Coenen, Zhengyong Jiang, and Jionglong Su. 2023. From deterministic to stochastic: an interpretable stochastic model-free reinforcement learning framework for portfolio optimization. *Applied Intelligence*, 53(12):15188–15203.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4971–4980. IEEE Computer Society.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 6558–6569. Association for Computational Linguistics.

Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L. Crowley. 2023. Accommodating missing modalities in time-continuous multimodal emotion recognition. In *International Conference on Affective Computing and Intelligent Interaction*, pages 1–8. IEEE.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, Lihuo He, and Xuemei Luo. 2023a. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259.

Ning Wang, Hui Cao, Jun Zhao, Ruilin Chen, Dapeng Yan, and Jie Zhang. 2023b. M2r2: Missing-modality robust emotion recognition framework with iterative data augmentation. *IEEE Transactions on Artificial Intelligence*, 4(5):1305–1316.

Yi Wang, Lanling Zeng, Jiaqi Zhang, and Yang Yang. 2026a. Zero-shot diffusive image restoration with consistency. *Signal Processing*, 248:110705.

Yiqi Wang, Zihao Yan, Jiaqi Zhang, Zhangkai Wu, Mingkai Zheng, Zequn Sun, Yanming Zhu, and Taotao Cai. 2026b. Map-graph: Provenance-aware shared memory for multi-agent workflows. *arXiv preprint arXiv:2608.10509*.

Yiqi Wang, Jiaqi Zhang, Taotao Cai, Zirui Liu, Qingqiang Sun, Zequn Sun, Zhangkai Wu, Manqing Dong, Mingkai Zheng, Xuefei Yin, and Yanming Zhu. 2026c. From agent traces to trust: A survey of evidence tracing and execution provenance in LLM agents. *CoRR*, abs/2606.04990.

Yuanzhi Wang, Zhen Cui, and Yong Li. 2023c. Distribution-consistent modal recovering for incomplete multimodal learning. In *IEEE/CVF International Conference on Computer Vision*, pages 21968–21977. IEEE.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023d. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*.

Jie Wen, Shijie Deng, Waikeung Wong, Guoqing Chao, Chao Huang, Lunke Fei, and Yong Xu. 2024. Diffusion-based missing-view generation with the application on incomplete multi-view clustering. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52762–52778. PMLR / OpenReview.net.

Junhao Xu, Zheng Liu, Mengqi Zhang, Guangyuan Dong, and Zhumin Chen. 2026. Lever can move

the earth: Towards adaptive semantic capacity balance for image-text retrieval. *IEEE Transactions on Multimedia*.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. 2024. Leveraging knowledge of modality experts for incomplete multimodal learning. In *ACM International Conference on Multimedia*, pages 438–446. ACM.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2025. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–27.

Yang Yang, Xi Zhang, Jiaqi Zhang, and Lanling Zeng. 2026. Parameterized image restoration with diffusion and gradient priors. *Knowledge-Based Systems*, 338:115488.

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. 2024. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *AAAI Conference on Artificial Intelligence*, pages 16416–16424. AAAI Press.

Caili Yu, Yiqi Wang, Jiaqi Zhang, Yiqun Duan, Mingkai Zheng, Zhangkai Wu, Kaize Shi, and Taotao Cai. 2026. From faulty memories to corrected actions: Dependency-guided rollback repair for memory-augmented agents. *arXiv preprint arXiv:2608.10502*.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *ACM International Conference on Multimedia*, pages 4400–4407. ACM.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *AAAI Conference on Artificial Intelligence*, pages 5634–5641. AAAI Press.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistic*, pages 2236–2246. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259.

Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. 2022. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2402–2415.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems*.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767. Association for Computational Linguistics.

Jiaqi Zhang, Xinjie Li, Guo Yang, Kang Yang, Wenjing Wang, and Yang Yang. 2026a. Intrinsic feature consistency learning network for generalizable medical image segmentation. In *International Conference on Intelligent Computing*, volume 16662 of *Lecture Notes in Computer Science*, pages 249–260. Springer.

Jiaqi Zhang, Zheng Pang, Rongrong Gao, Qiyuan Zhang, and Yang Yang. 2026b. Local epistemic uncertainty guided active sampling for plug-and-play diffusive image restoration. *arXiv preprint arXiv:2608.06981*.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2608–2618. Association for Computational Linguistics.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025a. Multimodal invariant sentiment representation learning. In *Findings of the Association for Computational Linguistics*, Findings of ACL, pages 14743–14755. Association for Computational Linguistics.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025b. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 22123–22138. Association for Computational Linguistics.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. 2025. Cmad: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *IEEE/CVF International Conference on Computer Vision*, pages 4626–4636.

Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. 2023. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

A Datasets and Evaluation Metrics

To verify the effectiveness of the proposed method, we conduct extensive experiments on three mainstream multimodal emotion recognition datasets. As shown in Table 6, the IEMOCAP dataset (Busso et al., 2008) contains five dyadic sessions of improvised or scripted dialogues involving ten actors, with two actors in each session. Following existing research settings (Liu et al., 2025, 2024), we evaluate the model on two utterance-level multi-class emotion classification tasks. The four-class setting contains 5,531 utterances from happy (with excited merged into happy), sad, neutral, and angry (Liu et al., 2025; Xu et al., 2024; Liu et al., 2024; Zuo et al., 2023; Zhao et al., 2021). The six-class setting contains 7,433 utterances from happy, sad, neutral, angry, excited, and frustrated (Liu et al., 2025; Mai et al., 2020; Lian et al., 2023; Majumder et al., 2019). For both IEMOCAP tasks, we adopt WA and UA as evaluation metrics.

As shown in Table 7, the CMU-MOSI dataset (Zadeh et al., 2016) contains 2,199 opinionated monologue video clips collected from YouTube, split into 1,284 training samples, 229 validation samples, and 686 test samples. Each utterance-level clip is annotated with a continuous emotion intensity score in $[-3, 3]$, ranging from highly negative to highly positive. As an extension of CMU-MOSI, the CMU-MOSEI dataset (Zadeh et al., 2018b) is a larger and more challenging corpus. It contains 22,856 video clips covering diverse topics, unconstrained environments, and different speakers, with 16,326 training samples, 1,871 validation samples, and 4,659 test samples. CMU-MOSEI uses the same continuous emotion intensity annotations in $[-3, 3]$. For CMU-MOSI and CMU-MOSEI, the evaluation metrics are selected according to the specific task setting. For binary emotion classification, we follow the standard non-zero binary classification protocol (Xu et al., 2024), where samples with scores greater than 0 are labeled as positive, samples with scores less than 0 are labeled as negative, and zero-score samples are excluded from binary evaluation. In the inter-modal missingness comparison and the noise robustness experiments, we report ACC and F1 for binary emotion classification. Specifically, CMU-MOSEI is evaluated with ACC and F1 in the main inter-modal missingness comparison, while both CMU-MOSI and CMU-MOSEI are evaluated with ACC and F1 in the noise robustness analysis.

Dataset		# Conversations	# Utterances
IEMOCAP Four-Class	Session1	28	1085
	Session2	30	1023
	Session3	32	1151
	Session4	30	1031
	Session5	31	1241
IEMOCAP Six-Class	Session1	28	1373
	Session2	30	1356
	Session3	32	1569
	Session4	30	1512
	Session5	31	1623

Table 6: Statistical information on IEMOCAP.

Dataset	Speaker	Video Clip	# Conversations			# Utterances		
			train	val	test	train	val	test
CMU-MOSI	89	2199	52	10	31	1284	229	686
CMU-MOSEI	1000	22856	2249	300	676	16326	1871	4659

Table 7: Statistical information on CMU-MOSI and CMU-MOSEI.

For the intra-modal missingness experiments on CMU-MOSI and CMU-MOSEI, following previous work (Zhu et al., 2025b), we report weighted F1 for binary emotion classification and MAE for emotion intensity prediction. MAE is computed between the predicted and gold emotion scores in $[-3, 3]$, where lower values indicate better regression performance.

B Implementation Details

B.1 Comparison Methods

For the inter-modal missingness setting, we compare PriMD with CPMNet (Zhang et al., 2022), MMIN (Zhao et al., 2021), GCNet (Lian et al., 2023), CIF-MMIN (Liu et al., 2024), MoMKE (Xu et al., 2024), CMAD (Zhuang et al., 2025), and HARDY-MER (Liu et al., 2025).

For the intra-modal missingness setting, we compare PriMD with P-RMF (Zhu et al., 2025b), MISA (Hazarika et al., 2020), MMIM (Han et al., 2021), TETFN (Wang et al., 2023a), ALMT (Zhang et al., 2023), TFR-Net (Yuan et al., 2021), and LNLN (Zhang et al., 2024). For these methods, we likewise do not re-implement or re-evaluate them. Instead, we mainly refer to the publicly reported results in LNLN (Zhang et al., 2024) and P-RMF (Zhu et al., 2025b).

Across the aforementioned experimental settings, we ensure that the pre-extracted multimodal features, data splits, and evaluation metrics are exactly the same as those used by the compared meth-

ods. To maintain consistency with prior work (Xu et al., 2024; Zhu et al., 2025b; Liu et al., 2025), we adopt the same protocol for fair comparison.

B.2 PriMD Implementation

To ensure the fairness of the experimental results, for the inter-modal missingness setting, we adopt the publicly available pre-extracted features provided by prior studies (Luo et al., 2023; Wang et al., 2023d; Zhao et al., 2021; Zuo et al., 2023), and evaluate the models under six fixed combinations of missing modalities. For the intra-modal missingness setting, we follow existing work (Zhang et al., 2024; Zhu et al., 2025b). We use their available pre-extracted features and gradually increase the missing rate from 0.0 to 0.9 with an increment of 0.1 to simulate missing data. Specifically, when the missing rate is 0.5, 50% of the information in each modality of the test data is randomly masked. PriMD is trained using PyTorch 2.7.0 with CUDA 12.8 on an NVIDIA RTX 5090 GPU with 32 GB of memory. For the IEMOCAP dataset, we perform five-fold cross-validation using the Leave-One-Session-Out strategy. This is a subject-independent evaluation protocol, since each session contains two unique actors who do not appear in any other session. For the CMU-MOSI and CMU-MOSEI datasets, we repeat each experiment five times and report the averaged results. The proposed framework is optimized with the AdamW optimizer. The batch size is set to 32, the initial learning rate is set to 1×10^{-4}, and the weight decay is set to 1×10^{-5}. The maximum number of training epochs is set to 100. Other detailed parameter settings of our method are reported in Table 8.

C Extended Main Experiments

C.1 Noise Robustness

To further evaluate the robustness of PriMD under input perturbations, following prior work (Gao et al., 2024), we add controlled noise with varying intensities to the multimodal features during training, validation, and testing. As shown in Table 9, we compare our method with MoMKE (Xu et al., 2024) and DiCMoR (Wang et al., 2023c) across multiple experimental settings. The performance of all methods decreases as the noise intensity increases. This indicates that noise corrupts the original modality representations and further increases the difficulty of incomplete multimodal emotion recognition. Compared with existing methods, PriMD maintains more stable average performance on CMU-MOSI, CMU-MOSEI, and the IEMOCAP Four-Class setting. For example, under strong noise with $\sigma = 20$, PriMD achieves average ACC/F1 scores of 53.67%/53.91% on MOSI and 61.93%/61.68% on MOSEI, as well as average WA/UA scores of 34.50%/34.14% on IEMOCAP, outperforming the compared methods overall. These results show that, when noise interference and modality missingness coexist, PriMD can reduce representation shifts caused by unreliable features and produce more robust predictions.

D Extended Ablation Study

We further supplement our analysis with comprehensive ablation studies on the core module and its internal mechanisms.

D.1 DPMC Components

In the ablation study of DPMC, we design four variants to analyze its effects. w/o DPMC removes the entire DPMC module and no longer uses the primitive memory bank. w/o Discrete Primitives removes the discretization process and directly uses continuous modality-specific representations for compensation. Shared Memory Bank allows the three modalities to share a single memory bank, instead of constructing separate modality-specific memory banks. Random Memory Bank uses a randomly initialized memory bank to verify the necessity of learnable primitives. As shown in Table 10, the full DPMC achieves the best performance under all missing-modality combinations, with average WA/UA scores of 77.43%/77.28%. Compared with w/o DPMC, the full model improves WA and UA by 1.48 and 1.52 percentage points, respectively, demonstrating the effectiveness of primitive memory for missing-modality compensation. The full model also maintains stable advantages over w/o Discrete Primitives and Shared Memory Bank, indicating that discrete primitive modeling and modality-specific memory banks help preserve modality-specific information. The performance of Random Memory Bank is overall lower than that of directly removing the DPMC module, suggesting that a randomly constructed memory bank introduces ineffective primitives, which instead interfere with missing-modality compensation and degrade model performance.

Table 8: Hyperparameter settings of PriMD.

Hyperparameter	IEMOCAP	CMU-MOSI	CMU-MOSEI
Optimizer	AdamW	AdamW	AdamW
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Weight decay	1×10^{-5}	1×10^{-5}	1×10^{-5}
Batch size	32	32	32
Epoch	100	100	100
d	128	128	128
d_r	64	64	64
Dropout rate	0.3	0.3	0.3
C	64	64	64
β	0.20	0.20	0.20
$K_{\max}$	16	16	16
γ	0.5	0.5	0.5
τ_{KD}	4.0	4.0	4.0
λ_{Align}	0.1	0.1	0.1
λ_{Indep}	0.01	0.01	0.01
λ_{VQ}	0.1	0.1	0.1

Table 9: Performance comparison on MOSI, MOSEI, and IEMOCAP datasets under varying noise levels (σ). The best results are highlighted. A, T, and V denote the acoustic, textual, and visual modalities, respectively. Gray shading indicates the presence of a modality, while gray dashed lines indicate its absence.

Dataset	Method	σ	A T V		A T V		A T V		A T V		A T V		A T V		Avg.	
			ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)
MOSI	DiCMoR	5	53.24	47.90	70.58	70.51	56.02	49.02	72.87	72.99	55.09	54.72	69.97	69.75	62.96	60.82
	MoMKE	5	52.90	52.25	58.69	58.90	52.13	51.83	59.49	59.41	52.29	51.79	57.77	57.93	55.55	55.35
	PriMD	5	53.13	52.84	73.06	73.81	57.64	56.27	73.89	73.93	56.03	57.12	74.56	75.06	64.72	64.84
	DiCMoR	10	50.46	45.91	59.72	57.88	51.85	47.91	61.57	58.04	54.63	50.62	63.43	59.81	56.94	53.36
	MoMKE	10	51.68	51.69	54.42	54.69	53.96	54.18	54.27	54.33	53.81	54.05	54.57	54.78	53.79	53.95
	PriMD	10	52.36	52.47	59.44	60.02	54.69	55.34	62.15	62.43	55.04	55.27	62.19	63.08	57.65	58.10
	DiCMoR	20	48.32	44.69	50.00	47.67	47.87	45.36	50.46	46.80	42.68	38.71	50.15	48.34	48.25	45.26
	MoMKE	20	52.44	52.61	54.27	54.51	53.35	53.64	52.13	52.42	51.22	51.36	53.35	53.60	52.79	53.02
	PriMD	20	53.54	53.46	55.27	55.61	54.16	54.27	52.09	52.73	52.69	52.45	54.28	54.94	53.67	53.91
MOSEI	MoMKE	5	60.48	55.91	77.74	77.67	63.87	60.83	77.66	77.10	63.04	59.63	77.11	76.94	69.98	68.01
	PriMD	5	61.83	56.97	79.17	78.85	64.52	64.73	79.15	79.52	66.08	66.47	79.34	79.66	71.68	71.03
	MoMKE	10	60.37	55.01	62.25	60.66	61.89	55.25	60.48	59.90	60.76	56.46	61.28	61.44	61.17	58.12
	PriMD	10	61.13	60.45	64.25	65.03	63.22	62.36	64.25	65.07	60.54	61.83	67.24	67.51	63.44	63.71
	MoMKE	20	59.74	55.50	60.13	55.45	57.65	55.00	58.67	56.52	58.06	54.85	60.26	54.95	59.09	55.38
	PriMD	20	61.15	60.24	63.27	63.44	59.68	59.73	61.28	61.94	60.41	59.86	65.77	64.85	61.93	61.68
			WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)
IEMOCAP Four Class	MoMKE	5	32.75	29.01	48.99	49.48	42.83	37.65	50.28	50.84	42.57	37.75	52.45	51.92	44.98	42.78
	PriMD	5	31.04	30.59	55.83	55.96	44.21	41.85	53.85	55.62	48.11	40.25	59.63	60.81	48.78	47.51
	MoMKE	10	31.47	27.67	32.71	29.46	39.08	31.17	32.15	28.78	36.83	31.77	34.36	31.32	34.43	30.03
	PriMD	10	31.85	30.54	36.77	39.82	42.01	40.53	48.71	49.22	40.26	39.14	44.36	44.12	40.66	40.56
	MoMKE	20	31.11	28.27	31.19	28.03	31.37	28.35	30.69	28.10	31.16	28.44	31.28	28.34	31.13	28.26
	PriMD	20	31.40	30.03	33.64	32.29	34.55	35.23	35.11	34.85	34.71	35.27	37.60	37.18	34.50	34.14

Table 10: Ablation results for DPMC components. The best results are highlighted.

Dataset	Modules	A T V		A T V		A T V		A T V		A T V		A T V		Avg.	
		WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)
IEMOCAP Four Class	w/o DPMC	72.41	73.15	82.88	82.91	63.12	61.05	81.95	82.74	73.88	73.61	81.45	81.12	75.95	75.76
	w/o Discrete Primitives	72.66	72.93	83.54	83.72	63.41	61.06	82.61	81.82	73.89	73.74	81.94	81.73	76.34	75.83
	Shared Memory Bank	72.59	73.84	83.45	83.37	63.62	61.68	82.49	83.06	74.27	73.94	82.09	81.86	76.42	76.29
	Random Memory Bank	68.45	69.62	80.43	80.46	61.35	60.13	80.47	80.96	71.88	71.65	80.15	79.84	73.79	73.78
	DPMC	73.82	74.46	84.13	84.09	64.77	62.81	83.65	84.83	75.06	74.84	83.17	82.66	77.43	77.28

Table 11: Effect of fixed Top-K and dynamic retrieval on Four-Class IEMOCAP. The best results are highlighted.

Method	WA (%)	UA (%)
Fixed Top-K		
$K = 1$	75.84	75.62
$K = 2$	76.31	76.08
$K = 4$	76.86	76.65
$K = 8$	76.58	76.41
$K = 16$	76.12	75.94
Dynamic Retrieval		
$K_{\max} = 16$	**77.43**	**77.28**

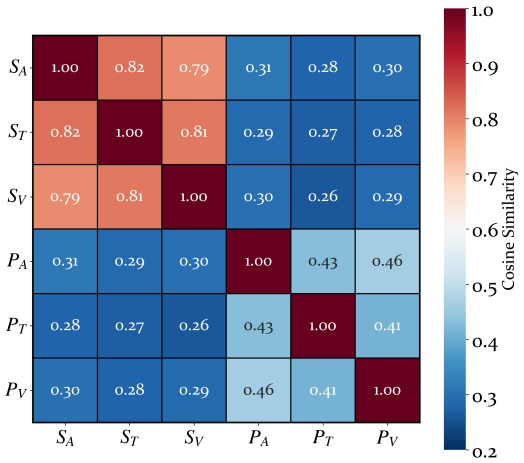


Figure 6: Visualization of shared-specific representation relations on Four-Class IEMOCAP.

D.2 DRAD Retrieval Analysis

D.2.1 Fixed Retrieval Size

To validate the effectiveness of the dynamic retrieval strategy in DRAD, we compare it with Top-K strategies using different fixed retrieval sizes. The fixed Top-K methods use the same number of primitives for all samples, with $K = 1, 2, 4, 8, 16$. In contrast, the dynamic retrieval method adaptively selects the primitives to retrieve according to sample characteristics, under the maximum retrieval budget of $K_{\max} = 16$. As shown in Table 11, the performance of fixed Top-K first increases and then decreases as K grows. The best fixed retrieval result is achieved when $K = 4$, with WA/UA of $76.86\%/76.65\%$. When K is too small, the number of retrieved primitives is insufficient, making it difficult to adequately compensate for missing-modality information. When K is too large, the additionally introduced primitives may contain redundant or weakly relevant information, which degrades representation quality (Lu et al., 2025; Qian et al., 2026b,a). By contrast, dynamic retrieval achieves the best performance, with 77.43% WA and 77.28% UA. It improves over the best fixed Top-K setting by 0.57 and 0.63 percentage points, respectively. This suggests that adaptively determining the retrieval size for different samples is more effective for missing-modality compensation.

D.2.2 Retrieval Cost Functions

As shown in Table 12, dynamic retrieval without a cost constraint improves WA/UA to 77.07%/76.91%, but uses 10.6 primitives on average. The linear cost reduces the expected retrieval size to 5.9 and yields 77.31%/77.16% WA/UA. The exponential cost further reduces the expected size to 5.2 and achieves the highest 77.43%/77.28% WA/UA. These results indicate that the exponential cost provides the best balance between recognition performance and primitive usage under the current setting.

D.3 Shared Alignment Objectives

As shown in Table 13, we compared cosine, InfoNCE, mutual-information, and L_2 alignment while keeping the remaining model architecture and training settings unchanged. Under the current naturally paired multimodal setting, L_2 alignment achieved the highest WA/UA of 77.43%/77.28%, slightly exceeding the 77.16%/76.98% obtained by cosine alignment, while InfoNCE and mutual-information alignment yielded lower performance. These results support the use of L_2 alignment in the current setting, but do not imply its universal superiority across datasets or pairing schemes.

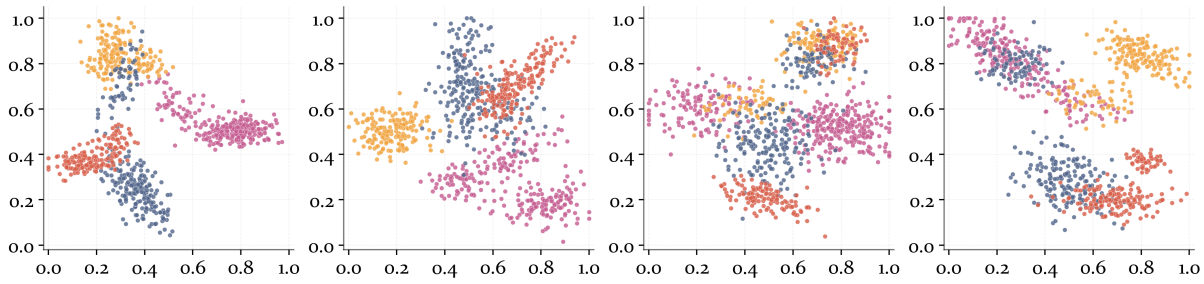
E Visualization Analysis

E.1 Shared-Specific Relations

To analyze the representation disentanglement effect of SSSD, we compute the average cosine similarity between S_A, S_T, S_V and P_A, P_T, P_V. As shown in Figure 6, the three shared representations have high similarity, indicating that the shared semantics across different modalities are well aligned. In contrast, the similarity between shared and specific representations is clearly lower, suggesting that they are effectively separated in the representation space. Meanwhile, the specific representations of different modalities maintain moderate similarity, which indicates that they still preserve

Table 12: Comparison of retrieval-cost functions on Four-Class IEMOCAP. The best results are highlighted.

Retrieval strategy	WA (%)	UA (%)	Expected r	Params	Inference (ms)
Fixed $K = 4$	76.86 ± 0.45	76.65 ± 0.48	4.0	2.327M	28.660
Dynamic, no cost	77.07 ± 0.46	76.91 ± 0.49	10.6	–	–
Dynamic, linear cost	77.31 ± 0.44	77.16 ± 0.48	5.9	–	–
Dynamic, exponential cost	77.43 ± 0.43	77.28 ± 0.47	5.2	2.336M	29.013



(a) Full (b) w/o SSSD (c) w/o DPMC (d) w/o DRAD

Figure 7: t-SNE visualization of fused representations on Four-Class IEMOCAP using only the text modality.

Table 13: Comparison of shared-alignment objectives on Four-Class IEMOCAP. The best results are highlighted.

Alignment	WA (%)	UA (%)
Cosine	77.16 ± 0.45	76.98 ± 0.49
InfoNCE	76.88 ± 0.50	76.71 ± 0.53
Mutual information	76.74 ± 0.52	76.56 ± 0.55
L_2	77.43 ± 0.43	77.28 ± 0.47

their modality-specific information and do not completely overlap. These results demonstrate that SSSD can achieve cross-modal alignment of shared semantics while retaining modality-specific information.

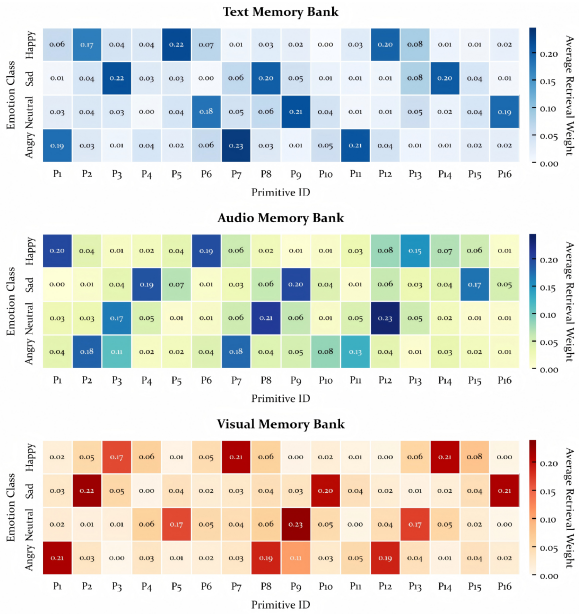


Figure 8: Visualization of class-level primitive retrieval patterns on audio, text, and visual memory banks.

E.2 Primitive Retrieval Patterns

To analyze the retrieval behavior of the primitive memory, we visualize the average primitive retrieval weights of different emotion categories on each modality-specific memory bank in the test set. As shown in Figure 8, the horizontal axis denotes the primitive ID, and the vertical axis denotes the emotion category. The results show that different emotion categories tend to activate different subsets of primitives within the same memory bank. For example, Happy, Sad, Neutral, and Angry correspond to distinct high-weight regions. This indicates that the retrieval process does not randomly select primitives, but is correlated with the emotion category of the sample. Meanwhile, the activation patterns of the audio, text, and visual memory banks are not identical, suggesting that the memory banks of different modalities learn their own modality-specific primitive distributions. In addition, the retrieval weights are not concentrated on a few fixed primitives. Instead, they form category-related usage patterns across multiple primitives. This suggests

that the primitive memory constructed by DPMC has certain diversity and can provide differentiated modality-specific information compensation for different emotion categories.

E.3 Fused Representations

To further analyze the effectiveness of different modules, we conduct t-SNE visualization on the IEMOCAP test set using only the text modality, comparing the classification representations of the full model and its ablated variants. As shown in Figure 7, we use pink, orange, gray, and yellow to denote Neutral, Happy, Sad, and Angry, respectively. The representations learned by the complete PriMD exhibit a clearer class structure, where samples from the same class are more compact and overlaps between different classes are reduced. In contrast, after removing SSSD, the sample distributions of different classes become more scattered, and clear mixing appears between some classes. After removing DPMC, the class boundaries become less distinct. After removing DRAD, the sample distribution becomes more entangled. These results indicate that the modules in PriMD jointly improve the class separability of emotion representations.