Rethinking Knowledge Distillation for Incomplete Multimodal Emotion Recognition: A Dynamic Approach

Anonymous Author(s)
Submission Id: 7528

Abstract

Multimodal Emotion Recognition (MER) is often hindered by missing modalities in real-world applications. While knowledge distillation offers a promising solution, existing methods tend to suffer from negative transfer due to inappropriate teacher selection, semantic mismatch between teacher and student, and conflicts in auxiliary gradients. From an information bottleneck perspective, a full-modality teacher is suboptimal for a modality-missing student, as excess heterogeneous signals introduce detrimental noise. To address these issues, we propose Dynamic Knowledge Distillation (DynKD), a unified self-distillation framework that operates across three dimensions. Structurally, DynKD restricts the teacher search space to limited extra modalities, using a Shapley-guided Probabilistic Routing Mechanism (SPRM) to dynamically select the optimal teacher configuration for each sample. At the feature level, a Feature Prior Mixer (FPM) aligns heterogeneous representations in a shared latent space, selectively injecting teacher priors to bridge semantic gaps without enforcing rigid imitation. At the optimization level, a Gradient Compatibility Rectification (GCR) module adjusts auxiliary gradients to amplify consistent updates and suppress conflicts. Extensive experiments demonstrate that DynKD significantly outperforms existing methods, achieving state-of-the-art robustness under diverse missing-modality conditions.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

Incomplete Multimodal Learning, Multimodal Emotion Recognition, Knowledge Distillation

1 Introduction

Multimodal Emotion Recognition (MER) aims to predict sentiment polarity and intensity by fusing complementary information from textual, acoustic and visual modalities [6, 12, 32]. However, most existing models [18, 19, 36, 44, 54–56] operate under the assumption that all modalities are available during training and inference. In real-world applications, unavoidable factors such as sensor constraints [42, 45], background noise [39, 58] and privacy concerns [1, 66] frequently lead to missing modalities. This data incompleteness severely degrades the performance of conventional models, making robust MER approaches indispensable for practical deployment.

Recently, numerous works have attempted to mitigate this issue, achieving significant progress [2, 14, 51, 64]. Existing methods can be broadly classified into three primary categories: modality generation, joint learning and Knowledge Distillation (KD). Modality generation methods rely on generative models [7, 11, 13, 24] to infer and reconstruct absent modalities based on available cross-modal information. However, these methods often yield uninterpretable and unstable results due to their sensitivity to network architectures [22, 23, 57]. Furthermore, high computational cost restricts their real-time applications [25, 51]. Alternatively, joint learning methods [21, 35, 66] map available modalities into a shared latent space via consistency constraints, enabling robust feature fusion for sentiment prediction. However, they frequently fail to preserve the modality-specific features, which are essential for maintaining discriminative information when other modalities are absent [50, 52]. To address these limitations, KD emerges as an effective paradigm that operates by distilling comprehensive multimodal representations from a teacher network trained with complete modalities to guide a student model constrained by missing inputs.

By distilling privileged information [28] from the teacher to the student, KD provides an efficient solution for incomplete MER [15–17, 68]. As shown in Figure 1, there are several distinct paradigms for implementing KD. The most intuitive approach involves first training the teacher models on complete modalities and subsequently using their logits [10], representations [37] or relations [34] to guide

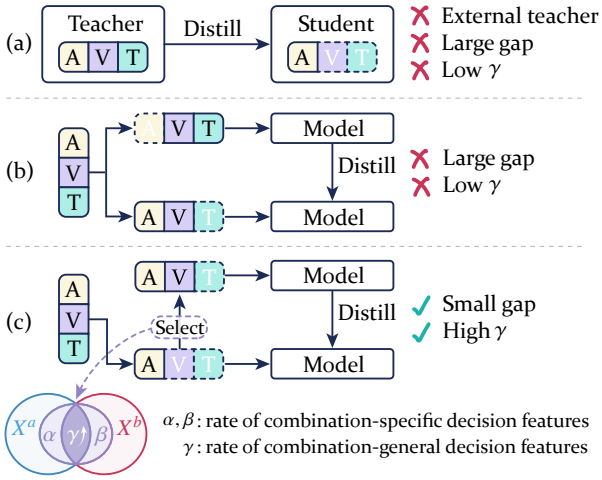

Figure 1: Architecture of different Knowledge Distillation (KD) paradigms for incomplete multimodal learning. A, V, and T denote acoustic, visual, and textual modalities, respectively. (a) Traditional KD, where an external teacher is trained on complete modalities; (b) conventional self-distillation, which transfers knowledge across arbitrary modality combinations; (c) our method, which selects a teacher with high modality overlap to ensure a small capacity gap and a higher proportion of shared decision-relevant features across modality combinations (γ).

the student model, which operates strictly under missing modalities. But these methods are limited by costly and teacher-dependent training [15], and the unrealistic assumption that missing modalities are strictly a test-time challenge [51]. To address these limitations, alternative approaches [15, 25] use self-distillation [63]. Instead of relying on auxiliary teachers, it enables bidirectional knowledge transfer by maintaining semantic consistency across different modality-missing versions within a single network.

However, existing KD methods for incomplete MER mainly focus on how to distill, while paying much less attention to which teacher modality combination provides the most suitable supervision. As in conventional KD, capacity gap [62] between teacher and student may also arise in incomplete multimodal distillation. Motivated by [53], we hypothesize that the effectiveness of incomplete multimodal distillation is closely related to the degree of modality overlap between the teacher and the student.

Table 1: Performance of distilling different teacher modality combinations to the audio-only student via logits. The best and second-best results are highlighted.

Dataset	Teacher	Student	ACC(%)	F1(%)
IEMOCAP (Six-class)	A T V	A	48.94	46.08
	A T	A	49.42	46.53
	T V	A	48.73	45.84
	T	A	49.31	46.14
IEMOCAP (Four-class)	A T V	A	65.62	64.51
	A T	A	66.38	65.62
	T V	A	65.04	64.03
	T	A	65.42	64.12

To validate this hypothesis, we present a pilot study in Table 1. On both the six-class and four-class settings of IEMOCAP, the audio-text teacher consistently outperforms the full-modality teacher when distilling to an audio-only student, despite using fewer modalities. In contrast, the text-visual teacher performs worse than the audio-text teacher and is even inferior to the text-only teacher in the six-class setting. These results suggest that, for incomplete multimodal distillation, teacher quality should be defined not only by informativeness, but also by compatibility with the student.

Based on the Information Bottleneck (IB) principle [40], using a full-modality teacher for a student with missing modalities is not ideal, as the extra signals often introduce noise. To address this issue, we propose Dynamic Knowledge Distillation (DynKD), a unified framework for incomplete MER. Instead of relying on a fixed full-modality teacher, DynKD adaptively selects a teacher modality combination that is better aligned with the student's observed inputs, achieving a better trade-off between informative guidance and noise suppression. Furthermore, as the utility of different modalities may vary across samples in the presence of real-world noise, static teacher assignment is often suboptimal. We therefore introduce a Shapley-guided Probabilistic Routing Mechanism (SPRM) to dynamically assess candidate teachers and route each sample to the most suitable modality configuration, providing adaptive and effective supervision.

To avoid the high cost of pre-training extra teachers, DynKD uses a self-distillation approach by randomly dropping modalities. However, this creates two main challenges during training: feature gaps and gradient conflicts. First, to handle feature gaps, we propose a Feature Prior Mixer (FPM), which aligns heterogeneous representations in a shared latent space and selectively incorporates teacher knowledge. This design enables the student to absorb compatible information without being forced to mimic a semantically mismatched teacher. Second, to fix gradient conflicts, we propose a Gradient Compatibility Rectification (GCR) module. It evaluates the consistency between auxiliary distillation signals and the primary task objective, boosting helpful updates and suppressing harmful ones. By adaptively choosing appropriate teachers and transferring only matching knowledge, our method reduces negative transfer and demonstrates superior robustness under missing-modality conditions compared to traditional distillation methods. Our key contributions are as follows:

- We propose a dynamic teacher selection method for incomplete MER based on the IB principle. Our method enables the model to select appropriate teacher modalities for each sample, thus leading to more effective supervision.
- We propose the DynKD framework, which features an SPRM for adaptive teacher routing, along with a FPM and GCR to resolve feature gaps and gradient conflicts during training.
- Extensive experiments demonstrate that DynKD significantly mitigates negative transfer and achieves state-of-the-art performance and robustness under various missing-modality conditions.

2 Related Work

2.1 Incomplete Multimodal Learning in MER

Missing modalities pose a significant challenge in practical MER. Extensive efforts have been devoted to this problem in recent years. The most direct approach is modality generation (or modality imputation). These methods attempt to reconstruct the absent modalities using information from the available ones. Early works simply used zero or average imputation, while subsequent deep learning methods utilize sophisticated generative models like VAE [13], GAN [7], and Diffusion models [11]. For instance, CRA [41] stacks cascaded residual autoencoders to simulate correlations between modalities. More recently, IMDer [47] leverages a score-based diffusion model to map random noise into the missing feature space. MPLMM [8] utilizes specific prompts to reconstruct missing inputs. Similarly, P-RMF [67] employs VAEs to project incomplete data into a latent Gaussian space, which generates a robust proxy modality to guide the reconstruction of missing information. Despite their effectiveness, these methods often incur high computational cost, which limits their practicality in real-time applications.

To alleviate the overhead of explicit generation, joint learning methods instead focus on learning robust cross-modal representations through shared latent spaces and consistency constraints. For example, MCTN [35] learns joint representations via cyclic translation on Seq2Seq models. MMIN [66] introduces a cross-modal

imagination module with cyclic consistency loss, and GCNet [21] utilizes graph neural networks for multimodal interaction. While effective, these methods typically emphasize shared joint representations and may overlook modality-specific features that are crucial for preserving discriminative information.

More recently, KD has emerged as an efficient and effective paradigm for incomplete multimodal learning. KD provides an efficient solution by distilling comprehensive multimodal representations from a teacher network trained with complete modalities to guide a student model constrained by missing inputs. Representative methods include UMDF [15], which proposes a unified self-distillation framework for uncertain missing modalities; CorrKD [17], which captures cross-sample and cross-category correlations via decoupled distillation; HRLF [16], which aligns teacher–student distributions through hierarchical mutual information maximization; and CMAD [68], which improves knowledge transfer via correlation-aware and modality-aware distillation. Existing KD-based methods for MER have shown promising results in incomplete multimodal learning. However, most of them still rely on a fixed teacher paradigm, where the teacher modality combination is predefined and shared across all samples, regardless of the missing-modality pattern or sample context.

2.2 Knowledge Distillation

Knowledge Distillation (KD) [10] transfers knowledge from a strong teacher to a weaker student. Early studies mainly focus on matching the teacher's output logits, while later works extend distillation to richer forms of supervision, such as intermediate representations [37] and relational structures [34]. Recent advances further improve the distillation objective itself. For example, DKD [65] decouples target and non-target logits for finer knowledge transfer, CTKD [20] uses a curriculum temperature to adapt the learning difficulty, and ABKD [43] introduces a generalized α-β-divergence to better balance confidence concentration and sample hardness.

Despite these advances, a fundamental challenge in KD is that a stronger teacher is not always a better teacher for the student. Prior studies have shown that an excessive teacher–student gap may hinder knowledge transfer, which motivates the use of teaching assistants [31], early-stopped teachers [4], and more calibration-robust distillation objectives [5]. SelKD [38] further suggests that only task-relevant knowledge should be distilled, since redundant supervision may overwhelm the student and even cause negative transfer. These findings indicate that effective distillation depends not only on what knowledge is transferred, but also on whether the teacher is suitable for the student.

This issue becomes more critical in incomplete multimodal learning. In this setting, the student operates under missing modalities, while the teacher usually has access to more complete inputs. Existing KD-based methods for MER have shown promising results by transferring logits, features, or relations from a modality-complete model to a modality-missing one. However, most of them mainly focus on what knowledge to distill, while paying less attention to the difficulty of distillation across different modality conditions. In particular, directly aligning teacher and student representations can cause semantic mismatch and optimization conflicts, because the teacher may depend on modality-specific information that the student cannot access. To address these issues, DynKD not only considers teacher selection, but also introduces dedicated designs to reduce the semantic gap and make optimization more stable, thereby helping reduce negative transfer.

3 Methodology

3.1 Overall Framework

As shown in Figure 2, the proposed DynKD framework consists of three core components: 1) SPRM, which dynamically selects the appropriate teacher modalities for each incomplete sample based on the sample-specific contributions of candidate modalities. 2) FPM, which transfers complementary semantic priors from the selected teacher to the student without enforcing rigid global alignment. 3) GCR, which calibrates auxiliary distillation gradients to better align with the primary task objective, thereby promoting stable optimization and alleviating gradient conflicts during training.

3.2 Shapley-guided Probabilistic Routing Mechanism

A full-modality teacher provides rich supervision, but may introduce modality-specific information that is not accessible to the student, leading to a large discrepancy between teacher and student. In contrast, a teacher built on the same modality subset S is fully compatible with the student, yet offers little additional cross-modal guidance. To balance complementary supervision and structural compatibility, we restrict the teacher search space to branches that extend S with only a limited number of unseen modalities.

Let $\bar{S}$ denote the set of unseen modalities. We define:

$$k = \min(K, |\bar{S}|), \tag{1}$$

where K is a predefined hyperparameter controlling the maximum allowable teacher–student gap. The candidate teacher space is then defined as:

$$C_k(S) = \left\{ S \cup A \;\middle|\; A \subseteq \bar{S},\ |A| = k \right\}. \tag{2}$$

By limiting the number of additional modalities, this design avoids overly distant teacher branches while still preserving useful complementary information. It also reduces the routing space and makes teacher assignment more efficient and stable.

A straightforward strategy would rank each unseen modality $m \in \bar{S}$ by its one-step marginal gain $V_x(S \cup \{m\}) - V_x(S)$. However, such a greedy criterion only measures the isolated effect of m and ignores that its utility may vary depending on which other unseen modalities are considered jointly. Since the Shapley value provides a principled way to quantify the contribution of each modality under different coalition contexts [9, 49], we adopt a Shapley formulation to obtain a sample-level estimate of modality contribution.

For a training sample (x, y) and an arbitrary modality subset $Q \subseteq M$, let $F_Q(x)$ denote the fused representation of Q. We define the sample-conditioned utility as:

$$V_x(Q) = -\mathcal{L}_{\text{task}}\big(g(F_Q(x)), y\big), \tag{3}$$

where $g(\cdot)$ is the prediction head and $\mathcal{L}_{\text{task}}$ is the task loss. In this way, the utility is directly tied to task performance on the current sample. We further define the marginal utility of adding modality m to the coalition $S \cup C$ as:

$$\Delta V_x(m; S, C) = V_x(S \cup C \cup \{m\}) - V_x(S \cup C), \tag{4}$$

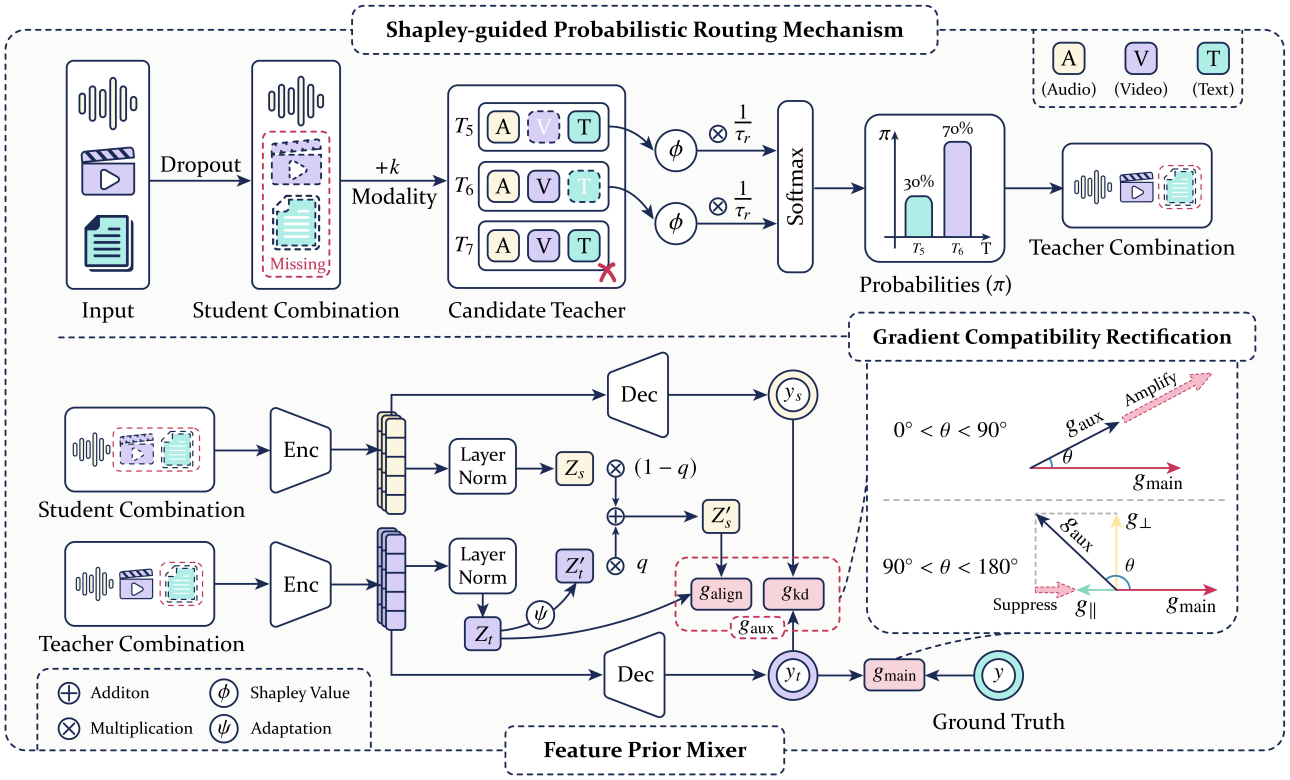

Figure 2: The overall architecture of the DynKD framework. It consists of three core components: 1) SPRM, which dynamically selects an optimal teacher branch for each incomplete sample based on Shapley values; 2) FPM, which softly transfers the teacher's cross-modal prior into the student's representation space without forcing exact imitation; and 3) GCR, which rectifies auxiliary distillation gradients dynamically to prevent interference with the main task optimization.

where $C \subseteq \bar{S} \setminus \{m\}$. Based on this quantity, the context-conditioned Shapley value of modality m is given by:

$$\phi_m(x,S) = \sum_{C \subseteq \bar{S}\setminus\{m\}} \frac{|C|!\,(|\bar{S}| - |C| - 1)!}{|\bar{S}|!}\,\Delta V_x(m;S,C). \tag{5}$$

Unlike simple marginal gains, $\phi_m(x,S)$ evaluates the expected contribution of m over all possible coalitions of other unseen modalities while keeping the subset S fixed. It therefore captures both the individual effect of m and its interactions with other modalities.

For a candidate teacher branch formed by an additional subset $A \subseteq \bar{S}$ with $|A| = k$, we define its coalition score as:

$$\Phi_A(x,S) = \sum_{m \in A} \phi_m(x,S). \tag{6}$$

We then convert these scores into a routing distribution over the candidate space:

$$\pi_A(x,S) = \frac{\exp\big(\Phi_A(x,S)/\tau_r\big)}{\sum\limits_{A' \subseteq \bar{S},\,|A'|=k} \exp\big(\Phi_{A'}(x,S)/\tau_r\big)}, \tag{7}$$

where τ_r is a temperature parameter controlling the sharpness of the routing policy. Rather than always selecting the highest-scoring branch deterministically, we perform stochastic routing to retain exploration among competitive candidates during training.

Finally, the target modality subset is sampled as:

$$\tilde{A} \sim \text{Categorical}\big(\pi_A(x,S)\big), \tag{8}$$

and the teacher branch assigned to the current sample is defined as:

$$T = S \cup \tilde{A}. \tag{9}$$

In this way, SPRM dynamically assigns a teacher according to its sample-dependent complementary value under the student's current observation pattern, rather than relying on a fixed rule based only on modality completeness.

3.3 Feature Prior Mixer

Although SPRM dynamically assigns a structurally compatible teacher T by sampling from $\pi_A(x,S)$, a semantic gap still remains between the student and the teacher branch with additional modalities. In particular, the teacher may contain extra modality-specific information that is not fully available to the student under the current incomplete input. Therefore, forcing the student representation to exactly match the selected teacher can be too rigid and may damage the student's own discriminative structure. To address this issue, we propose the FPM, which injects the teacher's additional semantic prior through a controlled adaptation and interpolation process, without requiring exact global alignment.

Given the fused student representation Z_S from the observable subset S and the fused teacher representation Z_T from the selected teacher branch, directly mixing them may introduce semantic noise, since they are derived from different modality combinations and are not fully aligned in feature space. To reduce this mismatch, we first apply a lightweight adaptation layer to the teacher representation:

$$Z'_T = \psi(Z_T), \tag{10}$$

where $\psi(\cdot)$ is a bias-free linear transformation parameterized by a weight matrix $W_\psi \in \mathbb{R}^{D\times D}$. This operation projects the teacher's cross-modal knowledge into a semantic subspace that is structurally compatible with the student. We adopt a lightweight linear adapter to perform only a mild semantic calibration, avoiding excessive transformation that could distort the teacher prior or introduce unnecessary parameters.

Next, we construct an enhanced student representation by interpolating the original student feature with this adapted teacher prior, governed by a constant interpolation factor $q \in (0, 1)$:

$$\hat{Z}_S = (1 - q)Z_S + qZ'_T. \tag{11}$$

Here, q explicitly controls the proportion of the teacher's prior injected into the student's context. This interpolation preserves the student's original discriminative basis while injecting only a controllable amount of complementary prior from the teacher.

Finally, the alignment loss for the selected teacher branch is formulated to pull this enhanced mixture toward the teacher's original feature space:

$$\mathcal{L}_{\text{align}} = \left\|\hat{Z}_S - \text{sg}(Z_T)\right\|_2^2, \tag{12}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. By minimizing this objective, the network effectively performs a cooperative reconstruction, i.e., instead of coercing Z_S to become identical to Z_T, the student learns to complement the adapted prior qZ'_T to approximate the full teacher representation. The external coefficient λ_{align} determines the overall influence of this feature-level distillation branch during training.

In addition, we impose response-level distillation to transfer the teacher's predictive behavior at the output level, where $y_s = g(\hat{Z}_S)$ and $y_t = \text{sg}(g(Z_T))$ denote the student and teacher outputs, respectively. For classification tasks, the distillation loss is:

$$\mathcal{L}_{\text{kd}} = \text{KL}(\text{softmax}(y_t/\tau_d) \,\|\, \text{softmax}(y_s/\tau_d)), \tag{13}$$

where τ_d denotes the distillation temperature. For regression tasks, we instead use:

$$\mathcal{L}_{\text{kd}} = \|y_s - y_t\|_2^2. \tag{14}$$

3.4 Gradient Compatibility Rectification

Despite the context-aware routing in SPRM and the adaptive interpolation in FPM, the auxiliary distillation objective may still conflict with the student's primary task during optimization. Since the unified model must accommodate diverse modality combinations, training naturally involves multiple optimization objectives [33, 48]. Consequently, teacher guidance may not always align with the student's own learning direction, leading to gradient interference. To address this issue, we introduce a GCR module.

Let Θ_s denote the shared parameter subset of the student network. We formulate the main-task gradient as:

$$g_{\text{main}} = \nabla_{\Theta_s}\mathcal{L}_{\text{task}}. \tag{15}$$

Similarly, the auxiliary gradients derived from the distillation branches are defined as:

$$g_{\text{kd}} = \nabla_{\Theta_s}\mathcal{L}_{\text{kd}}, \tag{16}$$

$$g_{\text{align}} = \nabla_{\Theta_s}\mathcal{L}_{\text{align}}. \tag{17}$$

For any auxiliary gradient $g_{\text{aux}} \in \{g_{\text{kd}}, g_{\text{align}}\}$, we first measure its cosine similarity with the main gradient to evaluate their optimization compatibility:

$$s = \cos(g_{\text{main}}, g_{\text{aux}}) = \frac{g_{\text{main}}^\top g_{\text{aux}}}{\|g_{\text{main}}\|\|g_{\text{aux}}\|}. \tag{18}$$

Instead of applying uniform scalar gating, GCR adopts a directional-aware hybrid rectification strategy. When $s \geq 0$, indicating that the auxiliary task is consistent with the main task, we amplify its effect to encourage synergistic learning:

$$\hat{g}_{\text{aux}} = (1 + \alpha s)g_{\text{aux}}, \tag{19}$$

where $\alpha > 0$ is a hyperparameter controlling the degree of positive-gradient amplification. Conversely, when $s < 0$, the auxiliary gradient directly interferes with the main task. Simply scaling down the entire gradient might discard valuable, non-conflicting structural knowledge. Therefore, we decompose g_{aux} into two orthogonal components with respect to g_{main}:

$$g_\parallel = \frac{g_{\text{aux}}^\top g_{\text{main}}}{\|g_{\text{main}}\|^2 + \varepsilon} g_{\text{main}}, \tag{20}$$

$$g_\perp = g_{\text{aux}} - g_\parallel, \tag{21}$$

where ε is a small constant for numerical stability. The parallel component $g_\parallel$ directly contributes to the conflict, while the orthogonal component $g_\perp$ does not explicitly oppose the main optimization direction. We then softly suppress the conflicting component while fully preserving the orthogonal one:

$$\hat{g}_{\text{aux}} = \rho g_\parallel + g_\perp, \tag{22}$$

where $\rho \in [0, 1]$ serves as a suppression coefficient. When $\rho = 1$, no rectification is applied; when $\rho = 0$, it degrades to a hard projection that completely removes the conflict. Intermediate values smoothly dampen the interference. After rectifying all auxiliary gradients, the final update direction for the shared parameters is:

$$g_{\text{total}} = g_{\text{main}} + \lambda_{\text{kd}}\hat{g}_{\text{kd}} + \lambda_{\text{align}}\hat{g}_{\text{align}}, \tag{23}$$

where λ_{kd} and λ_{align} balance the overall contributions of the response-level and feature-level distillation branches, respectively.

In this way, GCR preserves useful auxiliary information while reducing optimization conflicts with the main task.

4 Experiment

4.1 Datasets and Evaluation Metrics

To validate the effectiveness of our proposed approach, we conduct extensive experiments on three benchmark datasets:

The IEMOCAP dataset [3] comprises 4,453 video clips. Following standard practice, we evaluate our method on four-class (happy, sad, neutral and angry) [26, 27, 52, 66, 69] and six-class emotion recognition tasks (happy, angry, sad, neutral, surprised, fearful) [21,

Table 2: Performance comparison with state-of-the-art methods on two benchmark datasets under various missing modality scenarios. "Avg." indicates the mean performance across all settings. The best and second-best results are highlighted.

Dataset	Method	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	CPMNet [61]	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 [66]	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 [21]	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 [26]	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 [52]	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 [27]	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
	DynKD (our)	73.97/74.20	84.00/84.26	65.45/63.22	84.96/85.21	74.72/74.45	84.46/84.59	77.93/77.66
IEMOCAP six-class	CPMNet [61]	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 [66]	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 [21]	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 [26]	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 [52]	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 [27]	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
	DynKD (our)	54.35/51.36	63.63/61.88	45.91/43.57	65.54/63.92	54.32/50.60	64.27/62.77	58.00/55.68
		ACC(%)/F1(%)	ACC(%)/F1(%)	ACC(%)/F1(%)	ACC(%)/F1(%)	ACC(%)/F1(%)	ACC(%)/F1(%)	ACC(%)/F1(%)
CMU MOSEI	CPMNet [61]	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 [66]	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 [21]	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 [26]	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 [52]	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 [68]	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 [27]	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
	DynKD (our)	73.86/73.82	87.41/87.31	68.85/66.34	87.61/87.86	72.92/72.11	87.09/87.33	79.62/79.13

27, 29, 30]. We measure performance in terms of Weighted Accuracy (WA) and Unweighted Accuracy (UA).

The CMU-MOSI [60] and CMU-MOSEI [59] datasets comprise 2,199 and 22,856 opinionated YouTube video clips, respectively. Both datasets feature utterance-level annotations with continuous sentiment scores ranging from -3 (strongly negative) to +3 (strongly positive). In alignment with existing literature [27, 52], we frame this as a binary classification problem where scores > 0 are defined as positive and scores < 0 as negative. We employ Accuracy (ACC) and F1-score (F1) to assess performance.

4.2 Implementation Details

To ensure a fair and consistent comparison, we adopt the publicly available pre-extracted features provided by prior studies [21, 27, 52]. The proposed framework is optimized using the Adam optimizer, with momentum parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the initial learning rate to 1×10^{-4} and apply a weight decay of 1×10^{-5}. To mitigate overfitting, a dropout rate of 0.5 is applied to the student branch, and the batch size is fixed at 32 across all experiments. The teacher expansion hyperparameter is fixed at $K = 1$, while the remaining hyperparameters are empirically set as follows: $\tau_r = 0.1$, $\tau_d = 2.0$, $q = 0.3$, $\lambda_{\text{align}} = 0.1$, $\alpha = 0.1$, $\rho = 0.3$, and $\lambda_{\text{kd}} = 0.2$. All models are implemented in PyTorch 2.7.0 (CUDA 12.8) and trained on a single NVIDIA GeForce RTX 5090 GPU. For the IEMOCAP dataset, evaluation is conducted using five-fold cross-validation following the leave-one-session-out protocol, whereas for the CMU-MOSI and CMU-MOSEI datasets, we report the average performance over five independent runs.

4.3 Comparison with State-of-the-Art Methods

As shown in Table 2, we compare DynKD with several representative state-of-the-art baselines [21, 26, 27, 52, 61, 66, 68] under diverse missing-modality settings. Overall, DynKD achieves the best average performance on both IEMOCAP benchmarks and CMU-MOSEI. On IEMOCAP four-class, DynKD reaches 77.93% WA and 77.66% UA, improving over the strongest baseline, HARDY-MER, by 2.37% and 2.24%, respectively. On the more challenging IEMOCAP six-class task, DynKD further obtains 58.00% WA and 55.68% UA, surpassing the previous best average results by 1.29% and 3.03%. On CMU-MOSEI, DynKD also achieves the best overall performance, with an average ACC/F1 of 79.62%/79.13%. These results show that the proposed dynamic distillation framework is effective, and that routing incomplete samples to more suitable teachers can improve the quality of knowledge transfer during distillation.

As shown in Figure 3, in addition to comparing different missing-modality settings, we further evaluate methods [21, 35, 46, 66, 67] under varying missing rates. As the missing rate increases, the performance of all methods gradually declines. Nevertheless, DynKD remains competitive across different missing rates and consistently achieves superior overall performance, especially on CMU-MOSEI. These results further demonstrate the robustness of DynKD under diverse modality-missing scenarios.

Table 3: Ablation results under six missing conditions on the IEMOCAP four-class task. We report the weighted accuracy (WA) and unweighted accuracy (UA) for each missing condition, as well as the average performance over all six conditions. The best and second-best results are highlighted.

Dataset	Method	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	DynKD (our)	73.97/74.20	84.00/84.26	65.45/63.22	84.96/85.21	74.72/74.45	84.46/84.59	77.93/77.66
	w/o SPRM	73.48/73.89	83.17/83.28	64.78/62.22	84.87/85.18	74.12/74.12	84.73/84.71	77.53/77.23
	w/o FPM	73.08/73.20	83.21/83.17	65.19/61.77	84.71/84.93	73.28/72.36	84.41/84.61	77.31/76.67
	w/o GCR	73.04/72.51	83.83/83.95	64.52/62.18	85.18/85.75	73.15/73.17	84.47/84.36	77.37/76.99

4.4 Ablation Study

To demonstrate the necessity and effectiveness of each core module in DynKD, we perform extensive ablation experiments evaluated under the IEMOCAP four-class setting:

1) w/o SPRM: To evaluate the impact of SPRM, we replace the context-aware routing distribution with a uniform random selection. Specifically, in the w/o SPRM setting, the student still distills from a teacher with structurally compatible modalities, but the additional modality subset $\tilde{A}$ is sampled uniformly from the candidate unseen subsets of $\bar{S}$, rather than based on their expected complementary contributions. As shown in Table 3, this uninformed routing strategy leads to a noticeable performance drop. This confirms our hypothesis that different modalities provide varying utility depending on the incomplete context, and adaptively routing to the most informative teacher is crucial for optimal knowledge transfer.

2) w/o FPM: To further assess the contribution of the Feature Prior Mixer, we remove both the adaptation layer and the interpolation mechanism, and directly align the student representation with the teacher feature using a Mean Squared Error loss. This variant leads to a clear performance drop, which suggests that rigid feature-level matching tends to distort the student's original discriminative structure. The result highlights the importance of FPM, which enables a softer and more compatible form of feature transfer.

3) w/o GCR: To examine the importance of alleviating optimization interference, we completely remove the GCR module. In this setting, the auxiliary distillation gradients are directly aggregated with the main task gradient without any directional-aware amplification or orthogonal conflict suppression. The results in Table 3 show that unrectified gradients diminish the overall training efficacy, thus demonstrating that GCR successfully prevents negative interference while fully exploiting synergistic supervision.

4) Hyperparameter sensitivity: We further study the sensitivity of DynKD to the key hyperparameters involved in feature-level transfer, including the interpolation factor q and the alignment weight λ_{align}. As shown in Figure 4, DynKD remains stable across a broad range of settings and achieves the best performance in a moderate parameter region. In general, suitable values help the student absorb useful complementary information from the teacher, while overly large values gradually degrade performance by introducing overly strong constraints on feature alignment. These results suggest that DynKD benefits from a balanced feature-level transfer scheme, which helps reduce the semantic gap without disturbing the student's own representation learning.

5) Different teacher combinations: To validate our IB-inspired hypothesis, we compare DynKD with several teacher configurations: (a) a Random-Modality Teacher, which randomly selects a teacher modality combination without restricting the number of added modalities; (b) a Complementary Teacher, which uses only the modalities absent from the student; (c) a Full-Modality Teacher, which adopts the standard KD setting with all modalities available; and (d) our SPRM, which dynamically chooses the most suitable teacher configuration for each sample by introducing a constrained number of complementary modalities. As shown in Table 4, the competing teacher choices generally lead to inferior performance. The Random-Modality Teacher performs poorly because it ignores the compatibility between teacher and student inputs. The Complementary Teacher also yields limited gains, as using only missing modalities creates a large semantic gap from the student. Although the Full-Modality Teacher provides richer information, it also introduces harmful signals from unobservable modalities. In contrast, SPRM achieves the best results by striking a better balance between informativeness and compatibility. These results verify the effectiveness of our dynamic teacher selection strategy.

Table 4: Performance of different teacher combination strategies on IEMOCAP. The best and second-best results are highlighted.

Dataset	Teacher	WA(%)	UA(%)
IEMOCAP (Four-class)	(a) Random-Modality Teacher	77.45	77.19
	(b) Complementary Teacher	77.28	76.83
	(c) Full-Modality Teacher	77.37	76.98
	(d) DynKD (our)	77.93	77.66
IEMOCAP (Six-class)	(a) Random-Modality Teacher	57.72	54.74
	(b) Complementary Teacher	57.43	54.45
	(c) Full-Modality Teacher	57.49	54.80
	(d) DynKD (our)	58.00	55.68

4.5 Visualization Analysis

Figure 5 illustrates the evolution of teacher selection probabilities during training under two representative incomplete settings. For the audio-only student, the routing probability gradually concentrates on the *at* branch rather than *av*, which indicates that text provides more effective complementary information than vision in this case. For the visual-only student, the *tv* branch is consistently

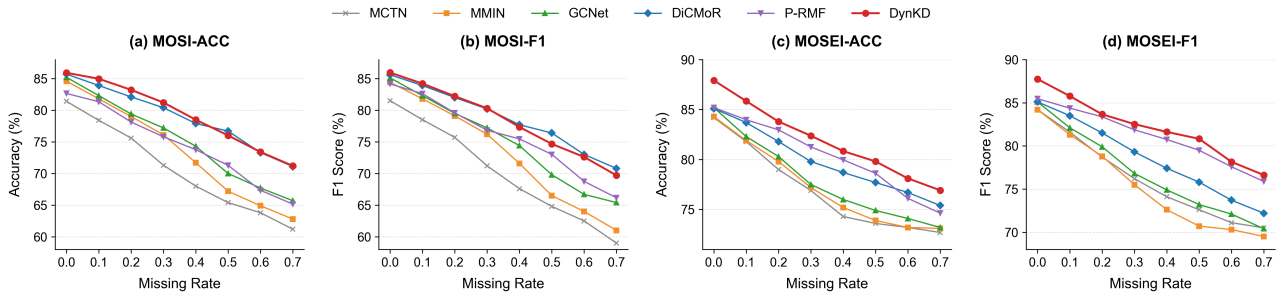

Figure 3: Performance curves of different methods under varying missing rates on the CMU-MOSI and CMU-MOSEI datasets. Subfigures (a) and (b) show the ACC and F1 results on MOSI, while (c) and (d) present the corresponding results on MOSEI.

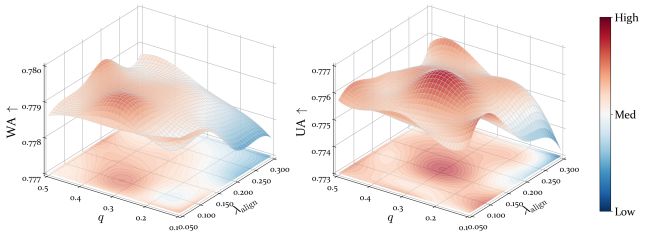

Figure 4: Hyperparameter sensitivity analysis on the IEMOCAP dataset. The figure presents the WA and UA under different combinations of the interpolation factor q and the feature alignment weight λ_{align}.

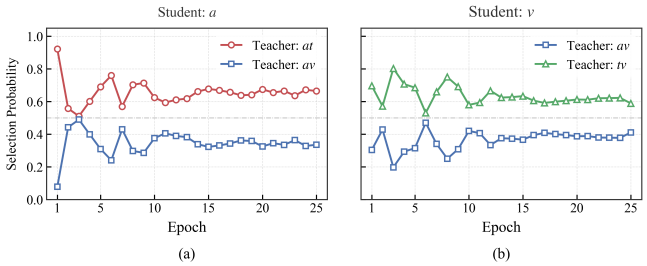

Figure 5: Teacher selection probabilities across training epochs under different student modality settings. For the audio-only student, the routing probability progressively favors the *at* branch over *av*; for the visual-only student, the *tv* branch remains dominant over *av*.

assigned a higher probability than *av*, which suggests that text is also the more compatible complementary modality for visual-only inputs. These results demonstrate that SPRM progressively learns sample-condition-aware routing preferences, instead of relying on a fixed teacher assignment strategy.

We further visualize the gradient angle distributions before and after correction in Figure 6 to examine how GCR influences the interaction between auxiliary objectives and the primary task. As shown in Figure 6(a), the KD gradient before correction exhibits a broadly spread distribution, indicating that its optimization direction varies substantially with respect to the task gradient. After

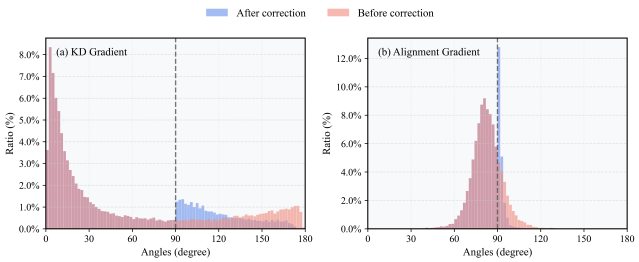

Figure 6: Distributions of gradient angles before and after gradient correction. (a) Distribution of the angles between the KD gradient and the task gradient. (b) Distribution of the angles between the alignment gradient and the task gradient. The dashed vertical line indicates the orthogonal direction at 90°. After correction, the auxiliary gradients are better aligned with the task gradient.

correction, the distribution is noticeably reshaped and becomes more concentrated around the conflict boundary, suggesting that unstable gradient directions are effectively moderated. A similar phenomenon can be observed in Figure 6(b) for the alignment gradient. Before correction, the angles are distributed over a relatively wider range, whereas after correction they become sharply concentrated around 90°, indicating that the corrected alignment gradient interacts with the task gradient in a more controlled manner. Overall, these results provide intuitive evidence that GCR can effectively regulate auxiliary gradients and stabilize their contribution to the optimization of the main task.

5 Conclusion

In this paper, we revisit knowledge distillation for incomplete MER from an IB perspective. We show that directly using a full-modality teacher often enlarges the semantic gap and introduces harmful noise, thus resulting in negative transfer. To address this, we propose DynKD, a unified self-distillation framework that performs compatibility-aware teacher routing via SPRM, soft prior transfer via FPM, and auxiliary gradient regulation via GCR. Extensive experiments on multiple benchmark datasets demonstrate that DynKD effectively reduces negative transfer and achieves state-of-the-art performance under diverse missing-modality settings.

References

[1] 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*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 991–1002.

[2] Reza Azad, Mohammad Dehghanmanshadi, Nika Khosravi, Julien Cohen-Adad, and Dorit Merhof. 2025. Addressing Missing Modality Challenges in MRI Images: A Comprehensive Review. *Computational Visual Media* 11, 1 (2025), 241–268.

[3] 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 (2008), 335–359.

[4] Jang Hyun Cho and Bharath Hariharan. 2019. On the Efficacy of Knowledge Distillation. In *IEEE/CVF International Conference on Computer Vision*. IEEE, 4793–4801.

[5] Wen-Shu Fan, Su Lu, Xin-Chun Li, De-Chuan Zhan, and Le Gan. 2024. Revisit the Essence of Distilling Knowledge through Calibration. In *International Conference on Machine Learning (Proceedings of Machine Learning Research)*. PMLR / OpenReview.net, 12882–12894.

[6] 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 (2023), 424–444.

[7] 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 (2014).

[8] 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*. Association for Computational Linguistics, 1726–1736.

[9] Kang He, Boyu Chen, Yuzhe Ding, Fei Li, Chong Teng, and Donghong Ji. 2026. PaSE: Prototype-aligned Calibration and Shapley-based Equilibrium for Multimodal Sentiment Analysis. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 30960–30968.

[10] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015).

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.

[12] Yingying Jiang, Wei Li, M. Shamim Hossain, Min Chen, Abdulhameed Alelaiwi, and Muneer H. Al-Hammadi. 2020. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion* 53 (2020), 209–221.

[13] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

[14] Lien P. Le, Thu Nguyen, Michael A. Riegler, Pål Halvorsen, and Binh T. Nguyen. 2025. Multimodal missing data in healthcare: A comprehensive review and future directions. *Computer Science Review* 56 (2025), 100720.

[15] Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. 2024. A Unified Self-Distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 10074–10082.

[16] 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. 2024. Toward Robust Incomplete Multimodal Sentiment Analysis via Hierarchical Representation Learning. In *Advances in Neural Information Processing Systems*.

[17] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024. Correlation-Decoupled Knowledge Distillation for Multimodal Sentiment Analysis with Incomplete Modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 12458–12468.

[18] Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled Multimodal Distilling for Emotion Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 6631–6640.

[19] Yong Li, Yuanzhi Wang, Yi Ding, Shiqing Zhang, Ke Lu, and Cuntai Guan. 2026. Decoupled Hierarchical Distillation for Multimodal Emotion Recognition. *CoRR* abs/2602.04260 (2026).

[20] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. 2023. Curriculum Temperature for Knowledge Distillation. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 1504–1512.

[21] 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 (2023), 8419–8432.

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

[23] 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*. IEEE, 211–221.

[24] 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.

[25] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. 2023. M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 1657–1665.

[26] 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 (2024), 1856–1873.

[27] 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*. ACM, 5755–5764.

[28] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. In *International Conference on Learning Representations*.

[29] 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*. AAAI Press, 164–172.

[30] 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*. AAAI Press, 6818–6825.

[31] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved Knowledge Distillation via Teacher Assistant. In *AAAI Conference on Artificial Intelligence*. AAAI Press, 5191–5198.

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

[33] Philip Novosad, Richard A. D. Carano, and Anitha Priya Krishnan. 2024. A Task-Conditional Mixture-of-Experts Model for Missing Modality Segmentation. In *Medical Image Computing and Computer Assisted Intervention (Lecture Notes in Computer Science)*. Springer, 34–43.

[34] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational Knowledge Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, 3967–3976.

[35] 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*. AAAI Press, 6892–6899.

[36] 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 (2025).

[37] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representations*.

[38] Liangliang Shi, Zhengyan Shi, and Junchi Yan. 2025. SelKD: Selective Knowledge Distillation via Optimal Transport Perspective. In *International Conference on Learning Representations*. OpenReview.net.

[39] Qiya Song, Jiajun Hu, Lin Xiao, Bin Sun, Xieping Gao, and Shutao Li. 2025. DiffCL: A Diffusion-Based Contrastive Learning Framework With Semantic Alignment for Multimodal Recommendations. *IEEE Transactions on Neural Networks and Learning Systems* 36, 10 (2025), 18587–18597.

[40] Naftali Tishby, Fernando C. N. Pereira, and William Bialek. 2000. The information bottleneck method. *CoRR* physics/0004057 (2000).

[41] 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*. IEEE Computer Society, 4971–4980.

[42] 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*. IEEE, 1–8.

[43] Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. 2025. ABKD: Pursuing a Proper Allocation of the Probability Mass in Knowledge Distillation via α-β-Divergence. In *International Conference on Machine Learning (Proceedings of Machine Learning Research)*. PMLR / OpenReview.net.

[44] Lan Wang, Junjie Peng, Cangzhi Zheng, Tong Zhao, and Li'an Zhu. 2024. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Information Processing & Management* 61, 2 (2024), 103675.

[45] Ning Wang, Hui Cao, Jun Zhao, Ruilin Chen, Dapeng Yan, and Jie Zhang. 2023. M2R2: Missing-Modality Robust Emotion Recognition Framework With Iterative Data Augmentation. *IEEE Transactions on Artificial Intelligence* 4, 5 (2023), 1305–1316.

[46] Yuanzhi Wang, Zhen Cui, and Yong Li. 2023. Distribution-Consistent Modal Recovering for Incomplete Multimodal Learning. In *IEEE/CVF International Conference on Computer Vision*. IEEE, 21968–21977.

[47] Yuanzhi Wang, Yong Li, and Zhen Cui. 2023. Incomplete Multimodality-Diffused Emotion Recognition. In *Advances in Neural Information Processing Systems*.

[48] Shicai Wei, Yang Luo, Yuji Wang, and Chunbo Luo. 2024. Robust Multimodal Learning via Representation Decoupling. In *European Conference on Computer Vision (Lecture Notes in Computer Science)*. Springer, 38–54.

[49] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. 2024. Enhancing Multimodal Cooperation via Sample-Level Modality Valuation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 27328–27337.

[50] 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 (Proceedings of Machine Learning Research, Vol. 235)*. PMLR / OpenReview.net, 52762–52778.

[51] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2026. Deep Multimodal Learning with Missing Modality: A Survey. *Transactions on Machine Learning Research* 2026 (2026).

[52] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *ACM International Conference on Multimedia*. ACM, 438–446.

[53] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. 2023. The Modality Focusing Hypothesis: Towards Understanding Crossmodal Knowledge Distillation. In *International Conference on Learning Representations*. OpenReview.net.

[54] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled Representation Learning for Multimodal Emotion Recognition. In *ACM International Conference on Multimedia*. ACM, 1642–1651.

[55] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. 2022. Emotion Recognition for Multiple Context Awareness. In *European Conference on Computer Vision (Lecture Notes in Computer Science)*. Springer, 144–162.

[56] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. 2022. Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences. In *ACM International Conference on Multimedia*. ACM, 1708–1717.

[57] 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*. AAAI Press, 16416–16424.

[58] 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*. ACM, 4400–4407.

[59] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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*. Association for Computational Linguistics, 2236–2246.

[60] 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 (2016).

[61] 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 (2022), 2402–2415.

[62] Chen Zhang, Qiuchi Li, Dawei Song, Zheyu Ye, Yan Gao, and Yao Hu. 2025. Towards the Law of Capacity Gap in Distilling Language Models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics, 22504–22528.

[63] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. In *IEEE/CVF International Conference on Computer Vision*. IEEE, 3712–3721.

[64] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and Changqing Zhang. 2024. Multimodal Fusion on Low-quality Data: A Comprehensive Survey. *CoRR* abs/2404.18947 (2024).

[65] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled Knowledge Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 11943–11952.

[66] 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*. Association for Computational Linguistics, 2608–2618.

[67] Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-Driven Robust Multimodal Sentiment Analysis with Incomplete Data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 22123–22138.

[68] 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*. 4626–4636.

[69] 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*. IEEE, 1–5.