State-Anchored Complete-View Distillation for Robust Conversational Multimodal Emotion Recognition

Zhaoyan Pan[1,*], **Xiangdong Li**[1,*], **Wenke Wu**[1,*], **Mengting Ma**[2],
Ye Lou[1], **Ji Zhou**[1], **Jiatong Pan**[1], **Wei Zhang**[1,2,†]
[1]Shcool of Software Technology, Zhejiang University
[2]Shcool of Computer Science and Technology, Zhejiang University
[1]{zhaoyanpan, xiangdong.li, wenkewu, ye_lou, jizhou, jiatongpan}@zju.edu.cn
[2]{mtma, ctszhangwei}@zju.edu.cn
[*]Equal contribution. [†]Corresponding author.

Abstract

Conversational multimodal emotion recognition (MER) requires reliable prediction when language, acoustic, or visual observations are missing or unreliable. Many missing-modality methods reconstruct absent inputs, yet such recovery can be non-unique in dialogue context, and nonverbal cues may conflict with the target utterance. To this end, we propose **CoRe-KD** (**Co**mplete-view **Re**ference-guided **K**nowledge **D**istillation), a state-anchored, conflict-regularized complete-view distillation framework for robust conversational MER. A complete-view teacher provides structured references, including prediction-level references, fused states, and modality-specific states. **Complete-view State Anchoring** (CSA) aligns incomplete-view student predictions and states with these references, while **Nonverbal Conflict Exposure** (NCE) trains on target-preserving nonverbal conflict views to reduce donor-label bias. Experiments on IEMOCAP and MELD, with CMU-MOSEI as a supplementary utterance-level check, show consistent gains under fixed- and random-missing protocols. Comprehensive ablation studies and further analyses support the role of CSA and the complementary effect of NCE.

1 Introduction

Conversational multimodal emotion recognition (MER) infers utterance-level emotions from dialogue context and language, acoustic, and visual observations (Busso et al., 2008; Poria et al., 2019). In practical scenarios, MER models often face incomplete observations, where any modality may be unavailable or partially observed, yielding insufficient evidence under fixed-missing or random-missing modality settings (Zhao et al., 2021a; Lian et al., 2023). Conversational nonverbal signals can also be loosely aligned with the target utterance due to speaker ambiguity, delayed reactions, or temporal mismatch (Tsai et al., 2019).

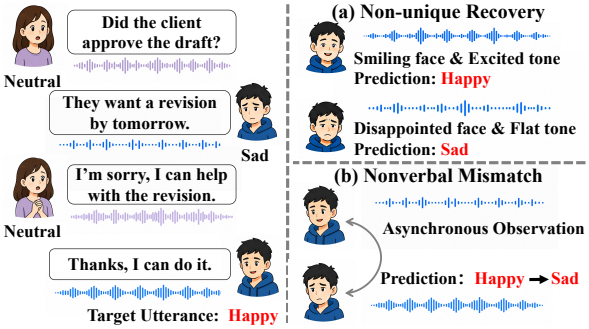


Figure 1: Panel (a) shows two issues: missing nonverbal cues lead to *non-unique recovery*, as the same utterance may admit multiple audio-visual completions; recovery or output-level distillation may leave *complete-view evidence underspecified* at fused and modality-specific levels. Panel (b) shows *target-inconsistent nonverbal evidence*, where asynchronous or mismatched cues may contradict the target utterance.

Recent robust MER methods have made progress in incomplete-input learning through reconstruction, distillation, and reliability-aware learning. Reconstruction-based methods estimate unavailable modality representations from observed inputs (Pham et al., 2019; Yuan et al., 2021; Zhao et al., 2021a; Guo et al., 2024; Zhang et al., 2024). Distillation- and alignment-based methods transfer complete-view predictions, representations, correlations, relations, or category-level knowledge to incomplete-view models (Wang et al., 2020; Lin and Hu, 2023; Li et al., 2023, 2024b; Ma et al., 2024; Zhuang et al., 2025). Reliability-, modality-, and expert-aware methods estimate confidence, uncertainty, quality, or modality-specific knowledge under partial modalities (Han et al., 2023; Xu et al., 2024; Bi et al., 2025; Dai et al., 2025; He et al., 2026; Miyoshi et al., 2026; Zhuang et al., 2026). These methods provide useful supervision for reconstructing unavailable information, transferring complete-view knowledge, or deciding how to weight and route observed modalities.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1663–1687
November 6-10, 2026 ©2026 Association for Computational Linguistics

However, as illustrated in Figure 1, incomplete and unreliable conversational observations raise two key challenges. *(i) Ambiguous recovery under missing observations*: Panel (a) shows that when nonverbal cues are absent, the observed language context does not uniquely determine the missing audio-visual evidence. The same utterance can correspond to multiple plausible nonverbal realizations, while the incomplete view still needs to retain the emotion-relevant information carried by the original complete observation. This makes the central question not only how to infer missing content, but also how to identify and preserve the complete-view evidence that is important at both the fused and modality-specific levels. *(ii) Target-inconsistent evidence under unreliable observations*: Panel (b) shows that available nonverbal cues are not always consistently aligned with the target utterance. Temporal asynchrony, weak correspondence, or mismatched audio-visual expressions can introduce evidence that is inconsistent with the target semantics and emotion, making direct reliance on all observed modalities unreliable. Together, these two challenges motivate using complete-view references to guide the preservation of fused and modality-specific evidence under missing observations, while explicitly exposing the model to target-inconsistent nonverbal evidence under unreliable observations.

To address these issues, we propose **CoRe-KD** (**Co**mplete-view **Re**ference-guided **K**nowledge **D**istillation), a state-anchored and conflict-regularized complete-view distillation framework for robust conversational MER. Different from methods that reconstruct missing inputs or mainly transfer teacher outputs, CoRe-KD treats the complete view as a structured task reference: predictive distributions guide incomplete-view decisions, fused Gaussian-inspired anchors preserve complete-view evidence, and modality-specific states serve as references for unavailable-modality-state alignment. Inspired by Product-of-Experts (PoE) modeling and multimodal PoE inference (Hinton, 2002; Wu and Goodman, 2018), these location-scale states provide precision-weighted fusion across different modality subsets. CoRe-KD then uses **Complete-view State Anchoring** (CSA) to align incomplete-view predictions, fused states, and unavailable-modality states with the complete-view teacher, avoiding the need to identify a unique raw completion (issue i) while explicitly preserving fused and modality-specific evidence (issue ii); it further uses **Nonverbal Conflict Exposure** (NCE) to train on target-preserving nonverbal conflict views, mitigating sensitivity to target-inconsistent nonverbal cues under controlled conflict views (issue iii). The main contributions of our paper can be summarized as follows:

- We propose **CoRe-KD**, a novel complete-view distillation framework that avoids input reconstruction by leveraging a complete-view teacher to provide three-level supervisory signals for robust conversational MER.
- We introduce two key components of CoRe-KD: **Complete-view State Anchoring** (CSA), which aligns incomplete-view predictions and teacher-provided anchors at both fused and modality-specific levels, and **Nonverbal Conflict Exposure** (NCE), which uses target-preserving nonverbal conflict views as a regularization signal to reduce the sensitivity to unreliable nonverbal observations.
- Extensive experiments on IEMOCAP and MELD demonstrate consistent improvements under fixed- and random-missing protocols, while CMU-MOSEI provides a supplementary utterance-level check.

2 Related Work

Multimodal Affective Computing. Multimodal affective computing models sentiment, emotion, and social-affective meaning from language, audio, vision, and conversational context (Busso et al., 2008; Poria et al., 2019). Classical studies explore multimodal fusion and cross-modal interaction through tensor fusion, low-rank fusion, memory fusion, Transformer-based alignment, and pre-trained language model integration (Zadeh et al., 2017; Liu et al., 2018; Zadeh et al., 2018; Tsai et al., 2019; Rahman et al., 2020). Recent studies further strengthen affective representation through correlation modeling, semantic constraints, context-dependent interaction, and fine-grained nonverbal affect modeling, including facial expression and action-unit generation (Xu et al., 2025; Pan et al., 2026; Li et al., 2026). Meanwhile, robust multimodal learning studies prediction under unavailable, degraded, or unreliable observations (Liang et al., 2021; Pham et al., 2019; Zhao et al., 2021a). Related sarcasm and misinformation studies further highlight semantic consistency, evidence reasoning, and robustness to noisy or incomplete cues (Wei

et al., 2024b,a; Yuan et al., 2025; Wei et al., 2025; Zhou et al., 2025b,a, 2026). Together, these studies motivate target-consistent affective evidence recovery in conversational MER, where nonverbal observations may be missing, weakly aligned, or inconsistent with the target utterance.

Robust Multimodal Emotion Recognition. Robust multimodal learning studies prediction under unavailable, degraded, or unreliable observations (Liang et al., 2021), while conversational MER focuses on utterance-level emotion recognition in dialogue, with IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019) as common benchmarks. Apart from complete-view distillation discussed below, existing methods mainly follow two lines. Reconstruction-based methods recover missing information, from early cross-modal translation or imagination methods in multimodal affective learning such as MCTN (Pham et al., 2019) and MMIN (Zhao et al., 2021a) to recent restoration-, diffusion-, prompting-, and proxy-based methods such as IMDer (Wang et al., 2023), MPLMM (Guo et al., 2024), and LNLN (Zhang et al., 2024). Recent reliability- and modality-aware robust multimodal methods further improve incomplete-view learning with modality experts, dynamic modality enhancement, cross-modal prompting, or uncertainty-regularized fusion, as represented by MoMKE (Xu et al., 2024), DREAM (Bi et al., 2025), ComP (He et al., 2026), and MCUR (Zhuang et al., 2026). These studies improve robustness when modalities are unavailable or incomplete. In dialogue, however, raw recovery can be non-unique, and nonverbal signals may be target-inconsistent.

Knowledge Distillation for Incomplete Multimodal Affective Computing. Knowledge distillation has been widely used in multimodal affective learning to transfer knowledge from complete, stronger, or teacher models to students under incomplete or heterogeneous inputs. Early incomplete-modality distillation uses complete-view teachers to guide missing-input students (Wang et al., 2020). Recent methods enrich the transferred signals with missing-modality supervision, decoupled factors, self-distillation, correlation knowledge, or modality-aware distillation, as in MissModal (Lin and Hu, 2023), DMD (Li et al., 2023), UMDF (Li et al., 2024a), Corr-KD (Li et al., 2024b), and CMAD (Zhuang et al., 2025). In conversational MER, SDT uses self-distillation for multimodal fusion (Ma et al., 2024), while TelME uses teacher-guided learning for modality interaction and dialogue-context modeling (Yun et al., 2024). These methods provide complete-view supervision at the output, representation, relation, correlation, or category level. Under condition-specific observations, fused and modality-specific complete-view states may remain implicit.

3 Preliminaries

3.1 Problem Formulation

Given a dialogue $\mathcal{D} = (u_1, \ldots, u_N)$, conversational MER predicts the emotion label $y_i \in \mathcal{Y}$ of target utterance u_i from context $\mathcal{D}_i$ and multimodal observations over $\mathcal{M} = \{\ell, \mathrm{a}, \mathrm{v}\}$. An observation condition $c \in \mathcal{C}$ specifies a modality-availability pattern. Under condition c, only modalities $\mathcal{A}_i^c \subseteq \mathcal{M}$ are accessible:

$$\begin{aligned} \mathbf{X}_i^c &= \{\mathbf{x}_i^{m,c} \mid m \in \mathcal{A}_i^c\}, \\ \hat{y}_i^c &= \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid \mathbf{X}_i^c, \mathcal{D}_i). \end{aligned} \tag{1}$$

The complete view has $\mathcal{A}_i^c = \mathcal{M}$; otherwise the view is incomplete. Target-preserving nonverbal conflict views are defined in Section 4.4.

3.2 Gaussian-inspired State Fusion

Gaussian-inspired State. We use a Gaussian-inspired state as a compact location-scale representation for modality fusion and teacher-reference matching (Wu and Goodman, 2018). It is used as a practical matching state rather than a calibrated probabilistic posterior. Let $\mathcal{G}(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$. For a generic modality representation $\mathbf{h}_{m,i}^c$, a state head produces a diagonal Gaussian-inspired state and its precision:

$$\begin{aligned} \left(q_{m,i}^c, \boldsymbol{\kappa}_{m,i}^c\right) &= \mathrm{StateHead}_m\left(\mathbf{h}_{m,i}^c\right), \\ q_{m,i}^c &= \mathcal{G}\left(\boldsymbol{\mu}_{m,i}^c, \boldsymbol{\sigma}_{m,i}^c\right). \end{aligned} \tag{2}$$

Here, $\boldsymbol{\kappa}_{m,i}^c = (\boldsymbol{\sigma}_{m,i}^c)^{-2}$ is the diagonal precision used for fusion. The implementation of $\mathrm{StateHead}_m$, including log-variance parameterization and clipping, is provided in Appendix C.2.

Product-of-Experts Fusion. Given the accessible modality set $\mathcal{A}_i^c$, available modality states are fused by a Product-of-Experts rule:

$$\begin{aligned} (\boldsymbol{\sigma}_i^c)^2 &= \left(\mathbf{1} + \sum_{m \in \mathcal{A}_i^c} \boldsymbol{\kappa}_{m,i}^c\right)^{-1}, \\ \boldsymbol{\mu}_i^c &= (\boldsymbol{\sigma}_i^c)^2 \odot \sum_{m \in \mathcal{A}_i^c} \boldsymbol{\kappa}_{m,i}^c \odot \boldsymbol{\mu}_{m,i}^c. \end{aligned} \tag{3}$$

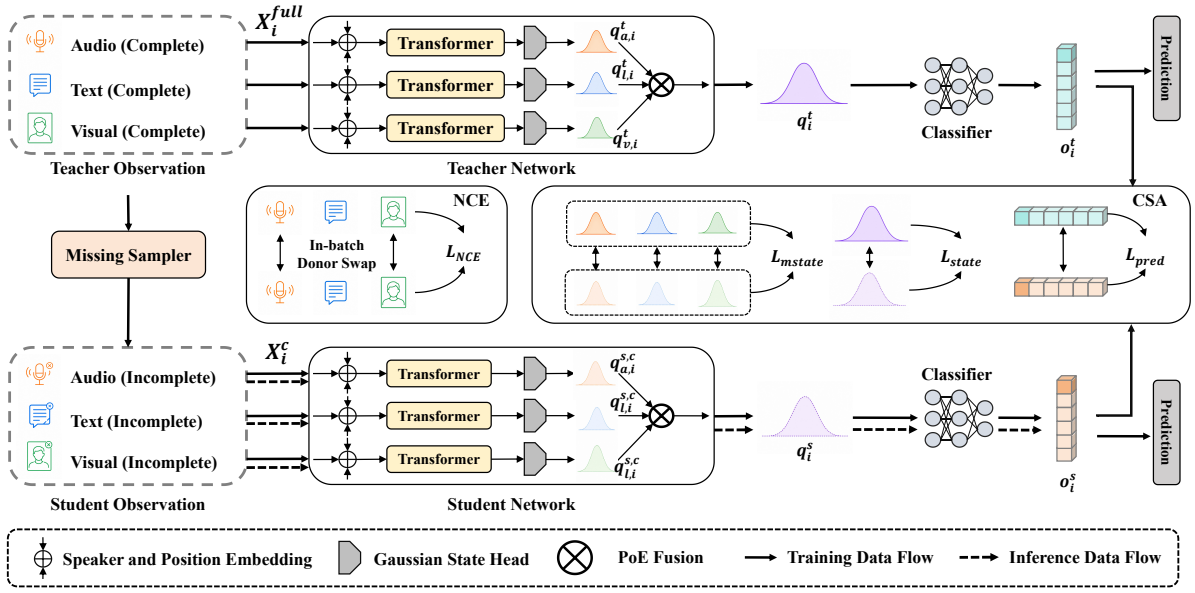


Figure 2: **Overall framework of CoRe-KD.** A frozen complete-view teacher provides prediction-, fused-state-, and modality-specific-state references for an incomplete-view student; CSA aligns these references to preserve complete-view evidence, NCE constructs target-preserving audio/visual conflict views via in-batch donor swaps, and only the student is retained at inference.

The $\mathbf{1}$ term corresponds to a zero-mean unit-precision prior and stabilizes fusion when few modalities are available. The fused state is denoted as $q_i^c = \mathcal{G}(\boldsymbol{\mu}_i^c, \boldsymbol{\sigma}_i^c)$.

4 Method

4.1 Overview

CoRe-KD follows a complete-to-incomplete training pipeline for robust conversational MER. A frozen complete-view teacher first encodes clean language, acoustic, and visual observations to produce prediction- and state-level references, while the student processes condition-specific views sampled from the same dialogue inputs. CSA anchors the student's predictions and states to the teacher reference, and NCE adds target-preserving nonverbal conflict views during training; only the student is retained for inference.

4.2 Reference State Estimation

For compact notation, we omit the dialogue context $\mathcal{D}_i$ in encoder equations. The generic representation $\mathbf{h}$ in Section 3.2 is instantiated by student encoder outputs: for each accessible modality $m \in \mathcal{A}_i^c$, the student encodes $\mathbf{x}_i^{m,c}$ into the target-utterance representation $\mathbf{h}_{m,i}^{s,c}$ and maps it to a state $q_{m,i}^{s,c}$ with precision $\boldsymbol{\kappa}_{m,i}^{s,c}$:

$$\begin{aligned} \mathbf{h}_{m,i}^{s,c} &= E_m^s\left(\mathbf{x}_i^{m,c}\right), \\ \left(q_{m,i}^{s,c}, \boldsymbol{\kappa}_{m,i}^{s,c}\right) &= \mathrm{StateHead}_m^s\left(\mathbf{h}_{m,i}^{s,c}\right). \end{aligned} \tag{4}$$

The accessible student-side states are fused by Eq. (3), yielding $q_i^{s,c} = \mathcal{G}(\boldsymbol{\mu}_i^{s,c}, \boldsymbol{\sigma}_i^{s,c})$.

On the clean complete view, the frozen teacher uses the same state interface and provides

$$\mathcal{T}_i = T(\mathbf{X}_i^{\mathrm{full}}, \mathcal{D}_i) = \left(\mathbf{o}_i^t, p_i^t, q_i^t, \{q_{m,i}^t\}_{m\in\mathcal{M}}\right), \tag{5}$$

where $\mathbf{X}_i^{\mathrm{full}}$ is the clean complete-view input; $\mathbf{o}_i^t$, $p_i^t = \mathrm{softmax}(\mathbf{o}_i^t)$, q_i^t, and $\{q_{m,i}^t\}_{m\in\mathcal{M}}$ denote the teacher logits, predictive distribution, fused state, and modality-specific states, respectively. The teacher implementation is detailed in Appendix C.2; teacher and student share the same state parameterization but not parameters.

Student-side logits and predictive distribution are $\mathbf{o}_i^{s,c} = C^s(\boldsymbol{\mu}_i^{s,c})$ and $p_i^{s,c} = \mathrm{softmax}(\mathbf{o}_i^{s,c})$. We use $\mathcal{K}_\tau$ for prediction matching and $\mathcal{D}_G$ for state matching:

$$\begin{aligned} \pi_\tau(\mathbf{o}) &= \mathrm{softmax}\left(\mathbf{o}/\tau\right), \\ \mathcal{K}_\tau(\mathbf{o}^t, \mathbf{o}^s) &= \tau^2 D_{\mathrm{KL}}\left(\pi_\tau(\mathbf{o}^t) \| \pi_\tau(\mathbf{o}^s)\right). \end{aligned} \tag{6}$$

For two diagonal states $q = \mathcal{G}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ and $q' = \mathcal{G}(\boldsymbol{\mu}', \boldsymbol{\sigma}')$, we define

$$\mathcal{D}_G(q, q') = \frac{1}{d}\left(\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}'\|_2^2\right), \tag{7}$$

where d is the state dimension. This squared 2-Wasserstein form provides the state distance for teacher-reference matching.

4.3 Complete-view State Anchoring

CSA anchors three student outputs to the complete-view teacher reference $\mathcal{T}_i$: predictions, fused states, and unavailable-modality states.

Prediction anchoring. CSA combines supervised classification with complete-view prediction distillation, so incomplete-view predictions remain guided by the complete-view decision:

$$\begin{aligned}\mathcal{L}_{\text{pred}}^{i,c} = \text{CrossEntropy}\left(y_i, \mathbf{o}_i^{s,c}\right) \\ + \lambda_{\text{kd}} \mathcal{K}_\tau\left(\mathbf{o}_i^t, \mathbf{o}_i^{s,c}\right).\end{aligned} \tag{8}$$

Fused-state anchoring. Beyond prediction-level supervision, CSA aligns the student fused state with the complete-view teacher state:

$$\mathcal{L}_{\text{state}}^{i,c} = \mathcal{D}_G\left(q_i^{s,c}, q_i^t\right). \tag{9}$$

Unavailable-modality-state anchoring. For unavailable modalities, CSA predicts teacher-side modality states instead of reconstructing raw inputs, providing modality-level references without requiring raw missing-input recovery. Let $\mathcal{R}_i^c = \mathcal{M} \setminus \mathcal{A}_i^c$ be the unavailable modality set, $\mathbf{a}_i^c$ the modality-availability indicator, and $\mathbf{z}_i^c = [\boldsymbol{\mu}_i^{s,c}; \boldsymbol{\sigma}_i^{s,c}; \mathbf{a}_i^c]$ the decoder input. For each $m \in \mathcal{R}_i^c$, a lightweight state decoder predicts the corresponding teacher-side Gaussian-inspired state:

$$\begin{aligned}\widehat{q}_{m,i}^{s,c} &= D_m\left(\mathbf{z}_i^c\right), \\ \widehat{q}_{m,i}^{s,c} &= \mathcal{G}\left(\widehat{\boldsymbol{\mu}}_{m,i}^{s,c}, \widehat{\boldsymbol{\sigma}}_{m,i}^{s,c}\right).\end{aligned} \tag{10}$$

The decoder follows the same Gaussian-inspired state parameterization as StateHead_m, but uses independent parameters.

For compactness, define the unavailable-modality state distance as $d_{m,i}^c = \mathcal{D}_G(\widehat{q}_{m,i}^{s,c}, q_{m,i}^t)$. The unavailable-modality-state anchoring loss is

$$\mathcal{L}_{\text{mstate}}^{i,c} = \begin{cases} \dfrac{1}{|\mathcal{R}_i^c|} \displaystyle\sum_{m \in \mathcal{R}_i^c} d_{m,i}^c, & |\mathcal{R}_i^c| > 0, \\ 0, & |\mathcal{R}_i^c| = 0. \end{cases} \tag{11}$$

The CSA training objective is

$$\begin{aligned}\mathcal{L}_{\text{CSA}} = \mathbb{E}_{i,c}\Big[\mathcal{L}_{\text{pred}}^{i,c} + \lambda_{\text{state}} \mathcal{L}_{\text{state}}^{i,c} \\ + \lambda_{\text{mstate}} \mathcal{L}_{\text{mstate}}^{i,c}\Big],\end{aligned} \tag{12}$$

where the condition c is sampled from the training observation conditions.

4.4 Nonverbal Conflict Exposure

NCE builds target-preserving conflict views as controlled perturbations by keeping the target language and label unchanged while replacing audio/visual cues with different-label donor cues, thereby exposing the student to misleading nonverbal observations while discouraging reliance on target-inconsistent nonverbal cues.

Conflict-view construction. For target sample i, a donor is a mini-batch sample with a different label:

$$\mathcal{J}_i = \{j \in \mathcal{I}_{\text{mb}} \mid y_j \neq y_i\}. \tag{13}$$

If $\mathcal{J}_i = \emptyset$, sample i is skipped for NCE; otherwise, $j \sim \text{Unif}(\mathcal{J}_i)$. Given donor j, we sample a nonverbal replacement subset $b \sim \text{Unif}(\mathcal{B})$, where $\mathcal{B} = \{\{\text{a}\}, \{\text{v}\}, \{\text{a}, \text{v}\}\}$, and construct

$$\widetilde{\mathbf{x}}_i^{m,b} = \begin{cases} \mathbf{x}_j^{m,\text{full}}, & m \in b, \\ \mathbf{x}_i^{m,\text{full}}, & m \notin b, \end{cases} \qquad m \in \mathcal{M}. \tag{14}$$

Because the language input and target-side supervision are kept fixed by construction, we use y_i as the conflict-view training label. The donor label y_j is used only for donor selection and margin interpretation. The student processes $\widetilde{\mathbf{X}}_i^b = \{\widetilde{\mathbf{x}}_i^{m,b}\}_{m \in \mathcal{M}}$ and outputs $\widetilde{\mathbf{o}}_i^{s,b}$.

Margin interpretation and NCE loss. Let $\Delta_{ij}^b = \widetilde{\mathbf{o}}_{i,y_i}^{s,b} - \widetilde{\mathbf{o}}_{i,y_j}^{s,b}$ be the target-versus-donor margin. Since $y_j \neq y_i$, the donor label is a non-target class in cross-entropy:

$$\text{CrossEntropy}\left(y_i, \widetilde{\mathbf{o}}_i^{s,b}\right) \geq \log\left(1 + \exp(-\Delta_{ij}^b)\right). \tag{15}$$

Thus, target-label cross-entropy implicitly discourages donor-label bias. The NCE loss is

$$\mathcal{L}_{\text{NCE}} = \mathbb{E}_{i,b,j}\left[\text{CrossEntropy}\left(y_i, \widetilde{\mathbf{o}}_i^{s,b}\right)\right], \tag{16}$$

where i is valid, $b \sim \text{Unif}(\mathcal{B})$, and $j \sim \text{Unif}(\mathcal{J}_i)$. NCE is used only during training.

Final objective. The objective of CoRe-KD is

$$\mathcal{L} = \mathcal{L}_{\text{CSA}} + \lambda_{\text{NCE}} \mathcal{L}_{\text{NCE}}. \tag{17}$$

During training, the teacher remains fixed. At inference time, only the student encoders, state heads, fusion module, and classifier are retained.

Testing Condition		{a}	{v}	{ℓ}	{ℓ, a}	{a, v}	{ℓ, v}	{ℓ, a, v}
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Six	IMDer	47.13/46.58	36.78/29.35	65.56/65.58	67.41/67.23	50.22/49.44	67.10/67.00	68.15/68.12
	Corr-KD	52.37/51.17	28.03/22.75	65.56/65.59	72.64/72.72	57.30/56.61	67.28/67.31	72.52/72.64
	LNLN	47.69/45.71	37.89/34.12	64.63/64.55	67.96/67.92	54.04/53.79	65.99/66.01	69.13/69.12
	MoMKE	48.80/47.06	35.24/29.84	65.19/65.36	68.70/68.79	51.88/51.45	67.53/67.68	69.93/70.06
	MCULoRA	53.17/51.26	38.32/30.82	65.93/65.94	69.25/69.23	60.32/59.70	67.65/67.15	70.92/70.85
	ComP	52.87/50.17	39.68/30.43	65.87/65.96	69.19/69.03	58.78/57.47	67.28/67.04	71.04/70.95
	CoRe-KD	**57.86/57.82**	**40.36/36.36**	**67.53/67.56**	**73.75/73.83**	**61.12/61.13**	**68.58/68.53**	**74.68/74.74**
Missing Rate		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Six	IMDer	67.10/66.99	65.80/65.55	64.82/64.71	62.66/62.45	59.52/59.39	57.98/57.70	54.53/54.09
	Corr-KD	69.99/70.01	68.21/68.24	67.10/67.17	65.31/65.42	63.46/63.07	61.18/60.64	54.47/53.65
	LNLN	68.21/68.26	67.22/67.16	65.13/65.31	63.65/63.50	61.31/61.02	60.26/59.86	54.78/54.44
	MoMKE	69.38/69.56	66.17/66.33	64.82/65.05	63.34/63.58	60.20/60.35	58.90/58.06	55.33/54.23
	MCULoRA	70.18/70.08	68.76/68.63	67.53/67.32	66.05/65.77	63.15/62.90	61.49/61.13	57.79/57.93
	ComP	70.24/70.18	68.45/68.02	67.47/66.83	65.68/65.35	64.39/64.20	61.98/61.84	57.98/56.42
	CoRe-KD	**73.55/73.53**	**72.83/72.88**	**71.37/71.49**	**71.35/71.59**	**70.12/70.19**	**69.25/69.29**	**68.45/68.30**

Testing Condition		{a}	{v}	{ℓ}	{ℓ, a}	{a, v}	{ℓ, v}	{ℓ, a, v}
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
MELD Seven	IMDer	35.03/20.41	35.03/20.41	66.54/65.36	66.54/65.37	35.03/20.41	66.53/65.36	66.54/65.36
	Corr-KD	44.87/33.23	41.45/30.38	65.56/65.39	66.00/65.78	43.69/33.39	65.61/65.49	65.95/65.81
	LNLN	48.28/36.88	48.04/31.50	66.26/65.21	66.62/65.75	48.64/38.14	66.25/65.27	66.69/65.86
	MoMKE	48.07/32.72	48.12/31.27	63.87/64.50	66.00/65.58	48.37/31.93	65.91/65.47	66.96/65.68
	MCULoRA	48.11/34.93	48.04/31.45	65.96/64.93	66.08/65.29	49.04/36.30	65.93/64.76	66.19/65.26
	ComP	49.23/36.42	48.12/31.27	66.07/64.57	66.73/65.35	49.25/37.15	66.05/64.62	66.74/65.44
	CoRe-KD	**49.77/41.08**	**48.35/32.24**	**68.28/67.25**	**68.47/67.35**	**50.19/38.59**	**68.66/67.97**	**68.85/67.41**
Missing Rate		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
MELD Seven	IMDer	64.41/62.73	62.63/60.41	60.92/57.97	58.96/55.09	57.23/52.32	55.33/49.12	53.43/45.32
	Corr-KD	64.09/63.77	61.85/61.36	59.31/58.51	56.58/55.48	54.28/52.76	52.08/50.15	49.23/46.73
	LNLN	64.80/63.65	62.67/61.12	60.98/58.81	59.08/56.10	57.66/53.78	55.61/50.48	53.52/46.39
	MoMKE	65.12/63.29	63.12/60.69	61.21/58.06	59.64/55.58	57.93/52.66	55.71/49.03	53.92/45.52
	MCULoRA	64.49/63.05	62.86/60.78	61.26/58.45	59.44/55.61	57.80/52.90	55.98/49.74	53.84/45.60
	ComP	64.76/62.99	63.01/60.69	61.49/58.50	59.75/55.87	58.01/53.15	56.18/50.04	54.07/46.07
	CoRe-KD	**66.36/65.56**	**64.18/63.28**	**61.84/60.70**	**60.11/57.61**	**58.81/56.55**	**56.32/52.78**	**54.18/48.42**

Table 1: Results on IEMOCAP-6 and MELD-7 under fixed-missing and random-missing settings.

5 Experiment

5.1 Dataset and Evaluation Metrics

Datasets. We evaluate CoRe-KD on IEMOCAP (Busso et al., 2008) under 4-/6-class protocols and MELD (Poria et al., 2019) under the standard 7-class protocol as the primary dialogue-context MER benchmarks. We further report CMU-MOSEI (Bagher Zadeh et al., 2018) results in Table 2 as a supplementary utterance-level generalization check, since MOSEI does not evaluate dialogue-context modeling.

Evaluation Protocols and Metrics. In prior IEMOCAP-based MER studies, accuracy-based metrics are commonly reported. However, MELD-7 has under-represented classes such as fear and disgust. We therefore report accuracy (Acc) and weighted-F1 (F1) for all evaluations, with F1 as the primary metric. We evaluate all methods under two missing-modality protocols.

Fixed Missing Protocol follows MMIN (Zhao et al., 2021a) and evaluates fixed available subsets from $\{a\}$, $\{v\}$, $\{l\}$, $\{l, a\}$, $\{a, v\}$, $\{l, v\}$, and $\{l, a, v\}$. The same subset is applied to all test samples, and results are reported per condition.

Random Missing Protocol follows IMDer (Wang et al., 2023) and samples per-sample missing patterns with ratios from 0.1 to 0.7, while retaining at least one modality available. All results are reported per missing ratio.

5.2 Implementation Details

CoRe-KD implementation. For CoRe-KD, the complete-view teacher is first trained on clean training data and then frozen. The student is trained with fixed-missing, random-missing, and NCE auxiliary views, without an additional clean-view-only student training stage. For random-missing views, the missing ratio is sampled from 0.1 to 0.7, while retaining at least one modality. We report mean performance over five random seeds under the same fixed- and random-missing evaluation protocols. Hyperparameters are provided in Appendix A.6.

Baseline implementation. For distillation baselines, we use the same clean complete-view teacher as CoRe-KD whenever applicable. For reproduced

Testing Condition		{a}	{v}	{ℓ}	{ℓ, a}	{a, v}	{ℓ, v}	{ℓ, a, v}
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
CMU-MOSEI	IMDer	68.55/67.14	63.07/63.27	85.55/85.41	85.25/85.04	67.36/67.16	84.78/84.67	85.22/85.10
	Corr-KD	63.79/51.12	66.57/61.41	84.37/84.37	84.73/84.73	68.38/64.76	83.13/83.17	82.72/82.76
	LNLN	71.46/69.01	65.30/60.34	86.19/86.26	86.54/86.61	71.66/69.96	86.19/86.24	86.16/86.20
	MoMKE	70.53/68.56	66.57/61.97	85.44/85.61	86.16/86.24	71.60/68.68	86.13/86.18	86.05/85.98
	MCULoRA	68.93/65.75	65.60/60.55	85.97/86.01	86.54/86.55	70.47/68.40	86.16/86.14	86.05/85.98
	ComP	69.07/67.50	66.29/62.67	86.19/86.21	86.02/86.02	71.05/69.69	**86.82**/86.77	85.86/85.80
	CoRe-KD	**72.95/72.39**	**67.01/67.24**	**86.49/86.49**	**86.98/86.98**	**72.81/72.67**	**86.82/86.85**	**87.12/87.16**
Missing Rate		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
CMU-MOSEI	IMDer	85.88/85.93	84.18/84.17	83.49/83.41	82.55/82.36	79.86/79.37	77.90/77.01	74.99/73.23
	Corr-KD	82.61/82.34	81.73/81.26	80.30/79.46	77.49/75.96	74.77/72.16	71.85/67.63	69.21/63.17
	LNLN	85.75/85.66	84.48/84.33	82.94/82.63	81.07/80.48	79.33/78.35	76.66/74.93	74.22/71.55
	MoMKE	86.24/86.21	84.73/84.64	83.43/83.19	81.43/80.91	79.36/78.38	76.33/74.52	72.81/69.49
	MCULoRA	85.72/85.71	84.67/84.60	83.49/83.33	82.09/81.80	80.77/80.29	78.65/77.72	76.58/74.76
	ComP	85.70/85.65	84.86/84.72	84.05/83.88	83.32/83.05	81.82/81.35	**81.28/80.58**	**79.18/78.02**
	CoRe-KD	**86.49/86.40**	**85.06/84.89**	**84.23/84.04**	**83.52/83.21**	**82.00/81.54**	81.10/80.42	78.92/77.77

Table 2: Supplementary generalization results on CMU-MOSEI under fixed-missing and random-missing settings. MOSEI is used as an utterance-level generalization check rather than an exhaustive SOTA comparison.

baselines, we use the same language, acoustic, and visual features as CoRe-KD whenever supported, and train/evaluate them under the same missing-modality protocols and seed settings. All experiments are run on four NVIDIA RTX A5000 GPUs. Detailed controls on data splits, feature inputs, checkpoint selection, and evaluation masks are provided in Appendix B.

5.3 Baselines

We compare CoRe-KD with representative robust multimodal emotion recognition methods under two missing-modality protocols: fixed missing and random missing. The baselines include reconstruction- or recovery-based methods, i.e., IMDer (Wang et al., 2023) and LNLN (Zhang et al., 2024); distillation- or expert-knowledge-transfer-based methods, i.e., Corr-KD (Li et al., 2024b) and MoMKE (Xu et al., 2024); and adaptation- or prompting-based methods, i.e., MCULoRA (Zhao et al., 2025) and ComP (He et al., 2026). Together, these methods cover recent representative solutions for incomplete multimodal observations.

5.4 Main Result

Main results. Table 1 reports the results on IEMOCAP-6 and MELD-7. Quantitatively, CoRe-KD consistently boosts missing-modality performance across fixed- and random-missing protocols. It achieves the best Acc/F1 in all conditions on IEMOCAP-6, obtains the best Acc/F1 in all fixed-missing conditions on MELD-7, and also achieves the best Acc/F1 at all random-missing rates on MELD-7. In average F1, CoRe-KD improves over the strongest baselines by 3.57/1.90 points under fixed missing and 6.22/2.08 points under random missing on IEMOCAP-6/MELD-7, with larger gains under severe missingness as modality availability drops. Qualitatively, these margins grow under heavy modality corruption and reveal the robustness of our distillation scheme. The improvements also hold in the complete-view condition {l, a, v}, suggesting that robust missing-view training does not come at the cost of full-modality performance.

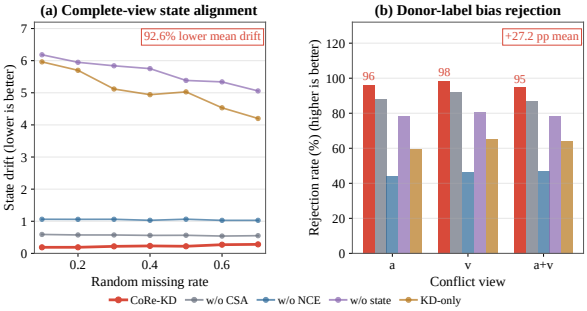


Figure 3: Mechanism analysis on IEMOCAP-6. (a) Lower state drift indicates better complete-view state alignment. (b) Higher rejection rate indicates better resistance to donor-label bias under nonverbal conflict.

Supplementary generalization. Table 2 reports CMU-MOSEI as a supplementary utterance-level generalization check. CoRe-KD obtains the best Acc/F1 in fixed-missing conditions and at random-missing rates from 0.1 to 0.5, while staying close to the strongest baseline at 0.6 and 0.7. In average F1, it gains 2.16 points under fixed missing and remains comparable under random missing, providing an additional check beyond the dialogue benchmarks.

5.5 Ablation Study

Table 3 reports ablations on IEMOCAP-6. CoRe-KD performs best across all ablation variants. Re-

Testing Condition		{a}	{v}	{ℓ}	{ℓ, a}	{a, v}	{ℓ, v}	{ℓ, a, v}
Dataset	Variant	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Six	w/o $\mathcal{L}_{\text{CSA}}$	46.27/45.29	33.33/32.45	63.52/63.61	72.15/72.41	50.28/48.92	65.43/65.66	72.89/73.12
	w/o $\mathcal{L}_{\text{NCE}}$	53.60/53.63	37.28/**36.73**	65.37/65.24	72.52/72.54	58.90/58.90	66.67/66.72	73.81/73.88
	w/o $\mathcal{L}_{\text{pred}}$	47.26/46.31	32.47/29.44	64.70/64.85	72.21/72.27	50.83/49.66	65.99/66.08	72.64/72.72
	w/o $\mathcal{L}_{\text{state}}$	54.90/54.15	36.29/32.29	67.22/67.15	73.26/73.22	58.90/58.14	68.21/68.13	73.88/73.83
	w/o $\mathcal{L}_{\text{mstate}}$	53.85/53.76	34.38/27.38	67.28/67.37	73.44/73.38	57.42/57.32	68.21/68.45	74.00/74.05
	CoRe-KD	**57.86/57.82**	**40.36**/36.36	**67.53/67.56**	**73.75/73.83**	**61.12/61.13**	**68.58/68.53**	**74.68/74.74**
Missing Rate		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Dataset	Variant	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Six	w/o $\mathcal{L}_{\text{CSA}}$	70.67/70.91	70.67/70.89	69.56/69.78	68.88/69.05	66.97/67.31	66.36/66.79	63.96/64.31
	w/o $\mathcal{L}_{\text{NCE}}$	72.15/72.21	71.10/71.20	69.81/69.92	68.58/68.71	68.02/68.27	67.90/68.21	66.48/66.70
	w/o $\mathcal{L}_{\text{pred}}$	70.43/70.52	70.30/70.38	69.99/70.07	69.44/69.39	68.76/68.90	66.91/67.08	63.83/63.71
	w/o $\mathcal{L}_{\text{state}}$	72.58/72.59	71.10/71.10	70.61/70.67	69.99/70.03	69.13/69.15	68.82/68.86	65.93/65.76
	w/o $\mathcal{L}_{\text{mstate}}$	71.78/71.88	70.67/70.77	70.12/70.23	70.61/70.75	69.62/69.80	68.64/68.61	66.30/66.21
	CoRe-KD	**73.55/73.53**	**72.83/72.88**	**71.37/71.49**	**71.35/71.59**	**70.12/70.19**	**69.25/69.29**	**68.45/68.30**

Table 3: Ablation of CoRe-KD on IEMOCAP Six under fixed-missing testing conditions and random-missing rates.

Testing Condition		$\{a\}$	$\{v\}$	$\{l\}$	$\{l, a\}$	$\{a, v\}$	$\{l, v\}$	$\{l, a, v\}$
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Four	MulT + Vanilla KD	45.97/43.39	41.71/31.05	76.73/76.74	77.62/77.78	56.84/56.99	74.07/73.79	73.51/73.54
	MulT + DKD	52.42/49.87	45.41/37.59	69.57/69.13	77.62/77.74	57.41/56.81	70.13/69.72	76.25/76.33
	MulT + Corr-KD	52.33/49.62	47.18/38.78	79.63/79.62	80.43/80.59	65.94/65.89	78.02/78.10	76.65/76.81
	MulT + CoRe-KD	**54.91/50.25**	**47.91/42.29**	**79.79/79.95**	**80.52/80.76**	**72.06/72.19**	**81.40/81.42**	**82.77/82.94**
	Δ	$+2.49/+0.38$	$+0.73/+3.51$	$+0.16/+0.33$	$+0.09/+0.17$	$+6.12/+6.30$	$+3.38/+3.32$	$+6.12/+6.13$
Testing Condition		$\{a\}$	$\{v\}$	$\{l\}$	$\{l, a\}$	$\{a, v\}$	$\{l, v\}$	$\{l, a, v\}$
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Four	MISA + Vanilla KD	44.61/37.79	44.85/37.00	63.20/61.68	71.42/71.40	48.23/43.39	63.28/59.84	71.42/71.57
	MISA + DKD	45.81/41.82	40.34/31.47	64.33/63.67	72.30/72.45	54.59/54.66	64.17/62.88	75.36/75.42
	MISA + Corr-KD	52.74/52.81	50.16/40.24	66.67/66.48	73.67/73.69	60.71/60.94	68.28/68.12	75.52/75.52
	MISA + CoRe-KD	**54.83/53.78**	**52.98/50.08**	**67.87/68.06**	**75.12/75.27**	**63.53/63.66**	**73.51/73.54**	**77.54/77.58**
	Δ	$+2.09/+0.97$	$+2.82/+9.84$	$+1.20/+1.58$	$+1.45/+1.58$	$+2.82/+2.72$	$+5.23/+5.42$	$+2.02/+2.06$

Table 4: Teacher-backbone generalization of CoRe-KD on IEMOCAP Four under fixed-missing testing conditions. Δ denotes the improvement of CoRe-KD over the strongest KD baseline under the same complete-view teacher.

moving $\mathcal{L}_{\text{CSA}}$ or $\mathcal{L}_{\text{pred}}$ causes the largest drops, indicating that complete-view anchoring and prediction supervision are the main contributors; removing $\mathcal{L}_{\text{state}}$, $\mathcal{L}_{\text{mstate}}$, or $\mathcal{L}_{\text{NCE}}$ also hurts in most cases, suggesting complementary gains from fused-state alignment, unavailable-modality-state alignment, and nonverbal conflict regularization. The performance gaps reveal that each objective fulfills a distinct role in stabilizing distillation under missing modalities. Appendix E controls for KD-only, hidden-state regression, raw reconstruction, deterministic state matching, and Gaussian CSA under the same backbone and training views, showing that the gains are not explained by prediction-level KD or generic auxiliary matching.

6 Further Analysis

To analyze the sources and teacher-backbone generality of CoRe-KD's gains, we conduct further studies around four questions. Appendix D gives lightweight theoretical notes on the design intuitions and scope.

Q1: Does CSA align with the complete-view teacher reference? We measure fused-state drift under the random-missing setting. For condition c, drift is $\mathbb{E}_i[(\|\mu_i^{s,c} - \mu_i^t\|_2^2 + \|\sigma_i^{s,c} - \sigma_i^t\|_2^2)/d]$, computed between the student fused state and the complete-view teacher state. As shown in Figure 3(a), CoRe-KD yields consistently lower drift than KD-only and ablated variants, with 92.6% lower mean drift than the average of compared variants. This indicates that CSA helps align the student state with the complete-view teacher reference beyond final prediction matching.

Q2: Does NCE reduce donor-label bias? We measure rejection rates under target-preserving nonverbal conflict views. For target i, donor j, and conflict type $b \in \mathcal{B}$, rejection is counted when $d_\mu(\widetilde{q}_i^{s,b}, q_i^t) < d_\mu(\widetilde{q}_i^{s,b}, q_j^t)$, where $d_\mu(q, q') = \|\boldsymbol{\mu}(q) - \boldsymbol{\mu}(q')\|_2$. As shown in Figure 3(b), CoRe-KD improves the mean rejection rate from 69.3% to 96.5% across $\{a\}$, $\{v\}$, and $\{a, v\}$ conflicts. This suggests that NCE helps preserve target-side evidence under donor-label nonverbal cues.

Q3: Are the gains tied to a specific teacher? To avoid drawing teacher-specific conclusions, we evaluate CoRe-KD with two alternative complete-view teachers, MulT (Tsai et al., 2019) and MISA (Hazarika et al., 2020), under the same fixed-

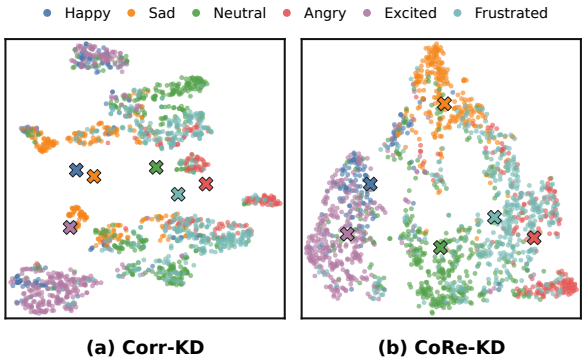


Figure 4: t-SNE visualization on IEMOCAP-6 under RMFM with missing rate 0.5. CoRe-KD yields more structured class distributions than Corr-KD.

missing protocol. As shown in Table 4, CoRe-KD improves across both teacher backbones, suggesting that the benefit is not tied to a particular complete-view teacher architecture and generalizes across different teacher designs.

Q4: Does structured distillation help beyond prediction transfer? Within each teacher backbone, all variants share the same frozen teacher, student architecture, input features, and missing-view schedule, differing only in the transferred knowledge: prediction-level KD transfers logits (Hinton et al., 2015), DKD decouples target/non-target knowledge (Zhao et al., 2022), Corr-KD transfers correlations, and CoRe-KD anchors fused and modality-specific Gaussian-inspired states. The consistent gains under MulT and MISA suggest that structured complete-view state anchoring provides useful supervision beyond prediction-level transfer for incomplete-observation MER, especially when modalities are partially missing and observation quality degrades.

Q5: Does CoRe-KD introduce substantial computational overhead? We report the training and inference costs in Table 5. Compared with the teacher, CoRe-KD increases student-training time from 0.14 h to 0.40 h due to the additional complete-view distillation and robustness objectives. Nevertheless, the absolute training cost remains modest. More importantly, the additional overhead at deployment is limited: inference latency increases only from 28.8 ms to 30.3 ms per batch (+5.2%), while the deployed model introduces 8.3M additional parameters (+9.4%). These results indicate that the main computational cost is concentrated during training, whereas the inference overhead remains small.

Item	Teacher	CoRe-KD	Change
Training time	0.14 h	0.40 h	+183.3%
Inference latency	28.8 ms	30.3 ms	+5.2%
Added parameters	–	8.3M	+9.4%

Table 5: Training and inference overhead of CoRe-KD.

7 Conclusion

In this paper, we proposed CoRe-KD, a structured complete-view distillation framework for robust conversational multimodal emotion recognition under unreliable observations. CoRe-KD uses Complete-view State Anchoring to align incomplete-view predictions and Gaussian-inspired states with complete-view teacher references, and Nonverbal Conflict Exposure to regularize target-preserving nonverbal conflict views. Experiments on IEMOCAP and MELD show improved robustness under fixed- and random-missing protocols, with mechanism analyses suggesting better complete-view state alignment and target preservation under controlled nonverbal conflict, while retaining a single student model at inference.

Limitations

CoRe-KD requires complete multimodal observations to train the frozen teacher, although inference only uses the student. Its NCE module relies on controlled target-preserving conflict views and may not cover all real-world multimodal misalignment or corruption patterns. In addition, the Gaussian-inspired state is used as a practical matching representation rather than a calibrated probabilistic posterior. Future work may study CoRe-KD under more diverse missing, noisy, and naturally conflicting multimodal conditions.

Ethics Statement

This work uses existing multimodal emotion recognition benchmarks and does not collect new human-subject data. For IEMOCAP, we obtained access via the official application procedure and used the data solely for permitted academic research, without redistribution or unauthorized use. MELD and CMU-MOSEI are publicly released research benchmarks. Since emotion recognition involves sensitive affective information, our results are intended for benchmark research only and should not be used for high-stakes decisions or user profiling without proper consent, privacy protection, and bias assessment.

References

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR.

AmirAli Bagher 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 Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.

Lanxin Bi, Yunqi Zhang, Luyi Wang, Yake Niu, and Hui Zhao. 2025. Two challenges, one solution: Robust multimodal learning through dynamic modality recognition and enhancement. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12855–12867, Suzhou, China. 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.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. 2025. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24507–24517.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736, Bangkok, Thailand. Association for Computational Linguistics.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2023. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Wen-Jue He, Xiaofeng Zhu, and Zheng Zhang. 2026. Cross-modal prompting for balanced incomplete multi-modal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 17463–17471.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10074–10082.

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

Xiangdong Li, Ye Lou, Ao Gao, Wei Zhang, and Siyang Song. 2026. Maugen: A unified diffusion approach for multi-identity facial expression and au label generation. *Preprint*, arXiv:2602.00583.

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

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.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. MultiBench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ronghao Lin and Haifeng Hu. 2023. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2024. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26:776–788.

Ryo Miyoshi, Mayu Otani, and Yuki Okafuji. 2026. Robust multimodal emotion recognition from incomplete modalities via query-based unimodal and cross-modal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4901–4911.

Zhaoyan Pan, Hengyang Zhou, Xiangdong Li, Yuning Wang, Ye Lou, Jiatong Pan, Ji Zhou, and Wei Zhang. 2026. Beyond isolated utterances: Cue-guided interaction for context-dependent conversational multimodal understanding. *Preprint*, arXiv:2604.25618.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proceedings of Interspeech 2019*, pages 3465–3469.

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 Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Qi Wang, Liang Zhan, Paul M. Thompson, and Jiayu Zhou. 2020. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838. Association for Computing Machinery.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, pages 17117–17128.

Yiwei Wei, Maomao Duan, Hengyang Zhou, Zhiyang Jia, Zengwei Gao, and Longbiao Wang. 2024a. Towards multimodal sarcasm detection via label-aware graph contrastive learning with back-translation augmentation. *Knowledge-Based Systems*, 300:112109.

Yiwei Wei, Shaozu Yuan, Hengyang Zhou, Longbiao Wang, Zhiling Yan, Ruosong Yang, and Meng Chen. 2024b. Gˆ 2sam: Graph-based global semantic awareness method for multimodal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9151–9159.

Yiwei Wei, Hengyang Zhou, Shaozu Yuan, Meng Chen, Haitao Shi, Zhiyang Jia, Longbiao Wang, and Xiaodong He. 2025. Deepmsd: Advancing multimodal sarcasm detection through knowledge-augmented graph reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, volume 31, pages 5580–5590.

Qinfu Xu, Yiwei Wei, Chunlei Wu, Leiquan Wang, Shaozu Yuan, Jie Wu, Jing Lu, and Hengyang Zhou. 2025. Towards multimodal sentiment analysis via hierarchical correlation modeling with semantic distribution constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21788–21796.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. 2024. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 438–446.

Shaozu Yuan, Yiwei Wei, Hengyang Zhou, Qinfu Xu, Meng Chen, and Xiaodong He. 2025. Enhancing semantic awareness by sentimental constraint with automatic outlier masking for multimodal sarcasm detection. *IEEE Transactions on Multimedia*.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407.

Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. TelME: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 82–95, Mexico City, Mexico. Association for Computational Linguistics.

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.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems*, volume 37, pages 55943–55974.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021a. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, Online. Association for Computational Linguistics.

Xinkui Zhao, Jinsong Shu, Yangyang Wu, Guanjie Cheng, Zihe Liu, Naibo Wang, Shuiguang Deng, Zhongle Xie, and Jianwei Yin. 2025. A robust incomplete multimodal low-rank adaptation approach for emotion recognition. *Preprint*, arXiv:2507.11202.

Zengqun Zhao, Qingshan Liu, and Shanmin Wang. 2021b. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556.

Hengyang Zhou, Yiwei Wei, Jian Yang, and Zhenyu Zhang. 2025a. Towards robust and relible multimodal misinformation recognition with incomplete modality. *arXiv preprint arXiv:2510.05839*.

Hengyang Zhou, Jinwu Yan, Yaqing Chen, Rongman Hong, Wenbo Zuo, and Keyan Jin. 2025b. Ldgnet: Llms debate-guided network for multimodal sarcasm detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Weilin Zhou, Zonghao Ying, Chunlei Meng, Jiahui Liu, Hengyang Zhou, Quanchen Zou, Deyue Zhang, Dongdong Yang, and Xiangzheng Zhang. 2026. Diver: Dynamic iterative visual evidence reasoning for multimodal fake news detection. *Preprint*, arXiv:2601.07178.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4626–4636.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. 2026. Modality-aware contrastive and uncertainty-regularized emotion recognition. *Preprint*, arXiv:2605.06245.

A Experimental Protocol

A.1 Datasets

IEMOCAP. IEMOCAP is used as one of our main dialogue-context MER benchmarks. We report both the commonly used four-class and six-class protocols. For IEMOCAP-4, we merge *happy* and *excited* and remove *frustrated*, resulting in four classes: *happy/excited*, *sad*, *neutral*, and *angry*. For IEMOCAP-6, we keep six emotion classes: *happy*, *sad*, *neutral*, *angry*, *excited*, and *frustrated*.

MELD. MELD is used as another main dialogue-context MER benchmark. It contains multi-party conversations with seven emotion labels: *anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness*, and *surprise*. Compared with IEMOCAP, MELD provides more diverse conversational interactions and a more imbalanced label distribution. In particular, *fear* and *disgust* are highly under-represented, making MELD a complementary testbed for evaluating robust conversational MER under imbalanced dialogue conditions.

CMU-MOSEI. CMU-MOSEI is included only as a supplementary utterance-level generalization check, rather than as a main dialogue-context benchmark. Following prior robust multimodal sentiment analysis studies, we use the non-zero sentiment setting. This setting helps examine whether CoRe-KD remains effective beyond dialogue-context MER, while the main claims are based on IEMOCAP and MELD.

Detailed dataset splits and the final training split are summarized in Table A.1; all final models are trained on the official training split.

Dataset	Dialogues	Utterances	Final Train
IEMOCAP-4	120 / – / 31	4,342 / – / 1,242	Train
IEMOCAP-6	120 / – / 31	5,810 / – / 1,623	Train
MELD-7	1,039 / 114 / 280	9,989 / 1,109 / 2,610	Train
CMU-MOSEI	– / – / –	12,786 / 1,438 / 3,634	Train

Table A.1: Dataset splits used in our experiments. Each split is reported as Train / Dev / Test, and the Final Train column indicates the split used for final model training. IEMOCAP and MELD are dialogue-context MER benchmarks; CMU-MOSEI is an utterance-level supplementary benchmark.

Dataset	Language	Acoustic	Visual	Feature Level
IEMOCAP	1024	1582	342	Utterance
MELD	768	768	768	Utterance
CMU-MOSEI	1024	512	1024	Utterance

Table A.2: Input feature dimensions used in our implementation. All features are pre-extracted and kept fixed during CoRe-KD training.

A.2 Feature Extraction

All experiments use pre-extracted utterance-level multimodal features. The feature extractors are kept fixed, and CoRe-KD only trains the dialogue encoder, modality projection layers, state heads, fusion module, and classifier. For dialogue-context datasets, utterance-level language, acoustic, and visual features are organized in dialogue order and processed as modality-specific sequences. Missing-modality experiments are conducted by applying availability masks to these pre-extracted modality features, without changing the underlying feature extraction pipeline.

For IEMOCAP, textual features are obtained from a RoBERTa-Large encoder (Liu et al., 2019) fine-tuned on conversation transcripts, using the last-layer `[CLS]` representation. Acoustic features are extracted with openSMILE (Eyben et al., 2010), and visual features are extracted with DenseNet (Huang et al., 2017) from facial frames. The resulting language, acoustic, and visual feature dimensions are 1024, 1582, and 342, respectively.

For MELD, textual features are extracted with a RoBERTa-based encoder (Liu et al., 2019) using speaker-aware dialogue context and a mask-prompt formulation, where the `<mask>` embedding is used as the utterance-level textual representation. Acoustic features are extracted from the target utterance speech segment with data2vec (Baevski et al., 2022), and visual features are extracted from target utterance video frames with TimeSformer (Bertasius et al., 2021). In the processed MELD features, all three modalities are stored as 768-dimensional utterance-level representations.

For CMU-MOSEI, audio, text, and visual embeddings are extracted with Wav2Vec (Schneider et al., 2019), DeBERTa (He et al., 2021), and MANet (Zhao et al., 2021b), respectively. The resulting language, acoustic, and visual feature dimensions are 1024, 512, and 1024, respectively. Each modality feature is first mapped to the model hidden dimension by a modality-specific projection layer before being passed to the intra-/inter-modal encoder.

A.3 Missing-Modality Evaluation Protocols

Fixed missing. The fixed-missing protocol evaluates robustness under deterministic modality availability. We consider seven available-modality subsets: $\{\mathrm{a}\}$, $\{\mathrm{v}\}$, $\{\ell\}$, $\{\ell, \mathrm{a}\}$, $\{\mathrm{a}, \mathrm{v}\}$, $\{\ell, \mathrm{v}\}$, and $\{\ell, \mathrm{a}, \mathrm{v}\}$, where ℓ, a, and v denote language, acoustic, and visual modalities. For each condition, the same modality subset is retained for all test samples, while the remaining modalities are treated as unavailable by the model. The complete-view condition $\{\ell, \mathrm{a}, \mathrm{v}\}$ is included to check whether robust training preserves full-modality performance rather than only improving incomplete-view cases.

Random missing. The random-missing protocol evaluates sample-wise stochastic modality loss. For each missing rate $r \in \{0.1, 0.2, \ldots, 0.7\}$, we independently generate modality-availability masks for test samples and ensure that at least one modality remains available. A higher r corresponds to a larger probability that each modality is unavailable. All methods are evaluated with the same generated masks under each seed, so performance differences are not caused by different missing-pattern samples. Results are reported separately for each missing rate to show how performance changes as modality incompleteness increases.

A.4 Training Views

During student training, we construct two types of incomplete views from each clean training sample. Fixed-missing views are generated by sampling one predefined non-empty available-modality subset, while random-missing views are generated by separately sampling a missing ratio from 0.1 to 0.7 and then drawing sample-wise modality-availability masks. All-missing random masks are rejected to ensure that at least one modality remains available. The student is optimized on these incomplete views, whereas the frozen complete-view teacher always receives the clean language, acoustic, and visual observations. The teacher provides prediction-level, fused-state, and modality-specific-state references, and no additional clean-view student-only branch is used.

A.5 Nonverbal Conflict View Construction

In NCE, we construct conflict views at training time by replacing nonverbal modalities while keeping the target language unchanged. For a target sample i, we sample a donor sample j from the same mini-batch with $y_j \neq y_i$. We then sample a replaced modality set $B \in \{\{\mathrm{a}\}, \{\mathrm{v}\}, \{\mathrm{a}, \mathrm{v}\}\}$. The conflict view keeps the target language feature and replaces the modalities in B with the corresponding donor features:

$$\widetilde{\mathbf{x}}_{m,i}^{B} = \begin{cases} \mathbf{x}_{m,j}, & m \in B, \\ \mathbf{x}_{m,i}, & m \notin B. \end{cases}$$

The supervision label remains the target label y_i. The donor label y_j is used only to ensure label mismatch during construction and to compute conflict-analysis diagnostics such as donor attraction or rejection rate. If no different-label donor is available in the mini-batch, the target sample is skipped for NCE.

This construction is used as a training-time stress regularizer for target-preserving misleading nonverbal evidence. It provides a controlled perturbation setting rather than an exhaustive simulation of all real conversational asynchrony patterns.

A.6 Optimization, Seeds, and Hardware

Optimization. All main experiments use the Adam optimizer with learning rate $1\mathrm{e}{-5}$, weight decay $1\mathrm{e}{-5}$, and batch size 16. IEMOCAP and MELD models are trained for 150 epochs. CMU-MOSEI models are trained for 100 epochs in the supplementary setting.

Loss weights. The KD temperature is set to $\tau = 2$. The loss weights are $\lambda_{\mathrm{kd}} = 1.0$, $\lambda_{\mathrm{state}} = 0.5$, $\lambda_{\mathrm{mstate}} = 0.5$, and $\lambda_{\mathrm{NCE}} = 1.0$. NCE views are sampled with probability 0.2. In the main implementation, NCE uses only the target-label conflict-view CE term; conflict-view KD and explicit margin losses are not used.

Training schedule and seeds. We train all models for the full schedule and apply the same predefined checkpoint-selection rule across compared methods. Unless otherwise specified, all IEMOCAP, MELD, and CMU-MOSEI results are reported as the mean over five random seeds. CMU-MOSEI is included as a supplementary utterance-level generalization check.

Hardware. All experiments are conducted on four NVIDIA RTX A5000 24GB GPUs.

A.7 Inference

At inference time, only the student model is used. The complete-view teacher, unavailable-state decoders, and NCE donor construction are used only during training and are discarded at inference.

Thus, CoRe-KD remains a single-student inference model under all missing-modality conditions.

B Baseline Fairness Control

B.1 Reproduction Setting

We compare CoRe-KD with representative incomplete-modality and robust MER baselines under the same fixed-missing and random-missing evaluation protocols. For reproduced methods, we retrain all models with the same data splits, checkpoint selection rule, and evaluation settings, and use the same feature inputs whenever feature replacement is supported. Whenever possible, we follow the original papers and released implementations; otherwise, we reimplement the method according to the original description and tune hyperparameters on the development split when available; for settings without a development split, we use fixed hyperparameters under the same protocol. Using shared extracted language, acoustic, and visual features whenever supported controls for feature-specific effects and makes the comparison focus on the multimodal learning strategy.

B.2 Baseline Categories

The compared baselines in Table **??** cover three representative families according to their primary modeling emphasis. IMDer and LNLN represent reconstruction- or recovery-oriented approaches. Corr-KD and MoMKE represent distillation- or expert-knowledge-transfer-based approaches. MCULoRA and ComP represent adaptation- and prompting-based approaches for incomplete multimodal learning. Together, these baselines cover recent major lines of missing-modality multimodal emotion recognition and provide a broad comparison for CoRe-KD.

B.3 Feature and Teacher Control

Baseline and feature control. To ensure controlled comparisons, reproduced baselines use the same language, acoustic, and visual features as CoRe-KD whenever feature replacement is supported. For methods whose official implementations are tightly coupled with their original feature preprocessing, we keep their original-feature setting to avoid feature-incompatibility artifacts while applying the same checkpoint selection rule. This reduces the risk of underestimating baselines due to feature incompatibility while keeping the main comparison as controlled as possible. Thus, performance differences are less likely to come from feature-specific advantages and more directly reflect the multimodal learning strategy.

Controlled distillation comparison. For distillation-based comparisons, we use the same clean complete-view teacher whenever applicable. This includes the vanilla prediction-level KD, DKD, and Corr-KD variants in the teacher-backbone generalization study. All distillation variants are trained with the same incomplete-view data and evaluated under the same fixed- and random-missing protocols as CoRe-KD. These comparisons test whether structured complete-view state anchoring provides useful supervision beyond prediction-level transfer in our robust MER setting. By sharing the same teacher source, the comparison isolates the effect of CoRe-KD's structured anchoring objective from the effect of teacher quality.

B.4 Training Schedule and Seeds

All compared methods are trained under the same main experimental protocol whenever applicable. We train each method for the full schedule with its recommended or tuned hyperparameters, and apply the same predefined checkpoint selection rule across fixed-missing and random-missing evaluations. Unless otherwise specified, results on IEMOCAP, MELD, and CMU-MOSEI are averaged over five random seeds. For each seed, all methods are evaluated using the same fixed-missing conditions and the same generated random-missing masks under the corresponding protocols. This unified training and evaluation protocol reduces variance from implementation choices and makes the reported differences more comparable across methods.

B.5 Stability Across Random Seeds

Seed-level stability. Table B.1 reports mean and standard deviation over five random seeds for the main fixed-missing and random-missing averages. Fixed Avg. is averaged over all fixed available-modality subsets, and RMFM Avg. is averaged over all evaluated random-missing rates. All methods use the same fixed-missing conditions and, for random-missing evaluations, the same generated masks under each seed, so the variance mainly reflects training stochasticity and random initialization.

Analysis. CoRe-KD achieves the highest average F1 across IEMOCAP-4, IEMOCAP-6, MELD-7,

Method	IEMOCAP-4		IEMOCAP-6		MELD-7		CMU-MOSEI	
	Fixed Avg.	RMFM Avg.	Fixed Avg.	RMFM Avg.	Fixed Avg.	RMFM Avg.	Fixed Avg.	RMFM Avg.
IMDer	68.35 ± 1.67	76.14 ± 1.02	56.19 ± 1.81	61.55 ± 0.68	46.10 ± 0.22	54.71 ± 0.20	76.83 ± 1.64	80.78 ± 2.38
LNLN	70.02 ± 1.11	77.29 ± 0.90	57.32 ± 0.17	62.79 ± 0.85	52.66 ± 0.36	55.76 ± 0.25	77.80 ± 0.06	79.70 ± 0.23
Corr-KD	74.07 ± 1.72	83.98 ± 0.85	58.40 ± 2.60	64.03 ± 1.30	51.35 ± 0.16	55.54 ± 0.08	73.19 ± 0.16	74.57 ± 0.73
MoMKE	71.85 ± 1.38	77.22 ± 2.05	57.18 ± 1.27	62.45 ± 1.93	51.02 ± 1.29	54.98 ± 0.38	77.60 ± 0.44	79.62 ± 0.32
MCULoRA	71.46 ± 2.47	79.89 ± 0.30	59.28 ± 0.83	64.82 ± 0.79	51.85 ± 0.11	55.16 ± 0.12	77.05 ± 0.66	81.17 ± 0.07
ComP	70.35 ± 1.20	79.86 ± 0.91	58.72 ± 1.38	64.69 ± 0.83	52.12 ± 0.28	55.33 ± 0.26	77.81 ± 0.34	82.46 ± 0.47
CoRe-KD	**80.85 ± 1.83**	**86.48 ± 0.71**	**62.85 ± 1.41**	**71.04 ± 0.76**	**54.56 ± 1.05**	**57.84 ± 1.35**	**79.97 ± 0.63**	**82.61 ± 0.23**

Table B.1: Seed-level stability under the main and supplementary missing-modality evaluations. Fixed Avg. is averaged over fixed-missing modality subsets, and RMFM Avg. is averaged over random-missing rates. Scores are reported as mean ± standard deviation over five random seeds. For IEMOCAP and MELD, the score is weighted F1; for CMU-MOSEI, we report the corresponding non-zero sentiment F1.

and CMU-MOSEI under both protocols. On the dialogue-context benchmarks, it improves over the strongest baseline by $6.78/2.50$, $3.57/6.22$, and $1.90/2.08$ points on IEMOCAP-4, IEMOCAP-6, and MELD-7 under Fixed/RMFM Avg., respectively. On CMU-MOSEI, CoRe-KD obtains the best fixed-missing average and remains comparable under random missing. Overall, the gains are larger than typical seed-level fluctuations, with standard deviations comparable to reproduced baselines, suggesting that the improvements are stable across seeds.

C Teacher Architecture

C.1 Overview

The complete-view teacher architecture and Gaussian-inspired state-head implementation are described below. For IEMOCAP and MELD, the hidden dimension is set to 1024; for CMU-MOSEI, it is set to 256. The Gaussian-inspired state dimension is $d = 256$. Each intra-/inter-modal Transformer uses one layer and 8 attention heads. The dropout rate is 0.5 for IEMOCAP and MELD, and 0.3 for CMU-MOSEI.

We do not use a sliding dialogue window. Instead, each dialogue-context sequence or utterance-level input sequence is padded and processed as a whole sequence. The maximum positional length is 512, and the maximum observed sequence lengths are 110, 33, and 87 for IEMOCAP, MELD, and CMU-MOSEI, respectively.

C.2 Complete-View Teacher

CoRe-KD uses a frozen complete-view teacher T to provide structured references for student training. The teacher is trained only with clean complete language, acoustic, and visual observations, and is fixed after convergence. In our implementation, T is instantiated as a Transformer-based multimodal dialogue encoder, following the intra-/inter-modal Transformer encoding design in prior multimodal sentiment and conversational emotion recognition models (Yuan et al., 2021; Ma et al., 2024). The teacher is used only as a complete-view reference model; CoRe-KD does not rely on teacher-specific reconstruction or self-distillation objectives.

For a dialogue context $\mathcal{D}_i$ around target utterance u_i, let

$$\mathbf{X}^t_{m,i} = \{\mathbf{x}^{m,t}_k \mid k \in \mathcal{C}_i\}, \qquad m \in \mathcal{M} = \{\ell, \mathrm{a}, \mathrm{v}\},$$

where $\mathcal{C}_i$ denotes the utterance-index set of the input sequence containing u_i. Each modality sequence is projected into a shared hidden space and augmented with positional and speaker/context embeddings:

$$\mathbf{Z}^t_{m,i} = \mathrm{Proj}^t_m\left(\mathbf{X}^t_{m,i}\right) + \mathbf{P}_i + \mathbf{S}_i. \tag{C.1}$$

Here, $\mathbf{P}_i$ and $\mathbf{S}_i$ denote positional and speaker/context embeddings; $\mathbf{S}_i$ is omitted for utterance-level datasets without speaker information.

The teacher first applies an intra-modal Transformer to encode contextual dynamics within each modality:

$$\mathbf{U}^t_{m,i} = \mathrm{IntraTrans}^t_m\left(\mathbf{Z}^t_{m,i}\right), \qquad m \in \mathcal{M}. \tag{C.2}$$

Here, $\mathrm{IntraTrans}^t_m(\cdot)$ is a self-attention Transformer encoder.

Following the common inter-modal Transformer design, modality m further attends to modality n. Let $\mathcal{I}^t_{mn}$ denote this inter-modal Transformer, where the first input provides queries and the second input provides keys and values:

$$\mathbf{V}^t_{mn,i} = \mathcal{I}^t_{mn}\left(\mathbf{U}^t_{m,i}, \mathbf{U}^t_{n,i}\right), \qquad n \neq m. \tag{C.3}$$

This produces cross-modal enhanced sequences $\{\mathbf{V}^t_{mn,i}\}_{n\neq m}$ for each target modality m.

For target utterance u_i, we select the target-position representations from the intra-modal and inter-modal sequences. Let $\mathrm{Sel}_i(\cdot)$ denote target-position selection:

$$\begin{aligned}\mathbf{r}^t_{m,i} &= \mathrm{Sel}_i\left(\mathbf{U}^t_{m,i}\right),\\ \mathbf{r}^t_{mn,i} &= \mathrm{Sel}_i\left(\mathbf{V}^t_{mn,i}\right).\end{aligned} \tag{C.4}$$

Inspired by hierarchical gated fusion, we combine the self-contextual and cross-modal representations by a lightweight gate. Let $\mathbf{c}^t_{m,i}$ be the concatenation of $\mathbf{r}^t_{m,i}$ and $\{\mathbf{r}^t_{mn,i}\}_{n\neq m}$. Then,

$$\begin{aligned}\mathbf{g}^t_{m,i} &= \mathrm{softmax}\left(\mathbf{W}^t_g\mathbf{c}^t_{m,i}+\mathbf{b}^t_g\right),\\ \mathbf{h}^t_{m,i} &= g^t_{m,0}\mathbf{r}^t_{m,i}+\sum_{n\neq m} g^t_{m,n}\mathbf{r}^t_{mn,i}.\end{aligned} \tag{C.5}$$

Here, $\mathbf{h}^t_{m,i}$ is the Transformer-enhanced teacher-side modality representation.

C.3 Gaussian-inspired State Head

Gaussian-inspired state heads are attached after these Transformer-enhanced modality representations. In the main text, $\mathrm{StateHead}_m$ denotes the abstract state-head interface. In this teacher implementation, it is instantiated by a lightweight Gaussian head GHead^t_m. For each modality, GHead^t_m outputs a $2d$-dimensional vector, which is split into a state location and a log-variance vector:

$$\left(\boldsymbol{\mu}^t_{m,i},\boldsymbol{\ell}^t_{m,i}\right)=\mathrm{Split}\left(\mathrm{GHead}^t_m\left(\mathbf{h}^t_{m,i}\right)\right), \tag{C.6}$$

where $\boldsymbol{\mu}^t_{m,i},\boldsymbol{\ell}^t_{m,i}\in\mathbb{R}^d$. The Gaussian head is implemented as a lightweight MLP with normalization, nonlinearity, dropout, and a final linear projection.

For numerical stability, we clamp the log-variance to $[\ell_{\min},\ell_{\max}]$, where $\ell_{\min}=-6$ and $\ell_{\max}=2$:

$$\boldsymbol{\ell}^t_{m,i}=\mathrm{clamp}_{[\ell_{\min},\ell_{\max}]}\left(\boldsymbol{\ell}^t_{m,i}\right). \tag{C.7}$$

The diagonal scale and precision are computed as

$$\begin{aligned}\boldsymbol{\sigma}^t_{m,i} &= \exp\left(\tfrac{1}{2}\boldsymbol{\ell}^t_{m,i}\right),\\ \boldsymbol{\kappa}^t_{m,i} &= \exp\left(-\boldsymbol{\ell}^t_{m,i}\right).\end{aligned} \tag{C.8}$$

The modality-specific teacher state is

$$q^t_{m,i}=\mathcal{G}\left(\boldsymbol{\mu}^t_{m,i},\boldsymbol{\sigma}^t_{m,i}\right). \tag{C.9}$$

C.4 Product-of-Experts Fusion

The complete-view fused teacher state is obtained by Product-of-Experts fusion over all modality states:

$$\begin{aligned}(\boldsymbol{\sigma}^t_i)^2 &= \left(\mathbf{1}+\sum_{m\in\mathcal{M}}\boldsymbol{\kappa}^t_{m,i}\right)^{-1},\\ \boldsymbol{\mu}^t_i &= (\boldsymbol{\sigma}^t_i)^2\odot\sum_{m\in\mathcal{M}}\boldsymbol{\kappa}^t_{m,i}\odot\boldsymbol{\mu}^t_{m,i},\\ q^t_i &= \mathcal{G}\left(\boldsymbol{\mu}^t_i,\boldsymbol{\sigma}^t_i\right).\end{aligned} \tag{C.10}$$

The $\mathbf{1}$ term corresponds to a unit Gaussian prior precision. It provides a weak fixed reference precision and is kept identical across teacher and student variants.

The teacher predicts emotions from the fused state location:

$$\mathbf{o}^t_i=C^t\left(\boldsymbol{\mu}^t_i\right),\qquad p^t_i=\mathrm{softmax}\left(\mathbf{o}^t_i\right). \tag{C.11}$$

It is trained with standard complete-view classification:

$$\mathcal{L}_T=\mathbb{E}_i\left[\mathrm{CE}\left(y_i,\mathbf{o}^t_i\right)\right]. \tag{C.12}$$

After training, all teacher parameters are frozen. During student training, the teacher always receives the clean complete view and provides the deterministic structured reference:

$$\mathcal{T}_i=\left(\mathbf{o}^t_i,p^t_i,q^t_i,\{q^t_{m,i}\}_{m\in\mathcal{M}}\right). \tag{C.13}$$

The student uses $\mathcal{T}_i$ for prediction anchoring, fused-state anchoring, and unavailable-modality-state anchoring.

C.5 Student State Heads and Unavailable-State Decoders

The teacher state heads, student state heads, and unavailable-state decoders use the same log-variance Gaussian-inspired state parameterization, but all parameters are independent. For an unavailable-state decoder D_m, the output is split into a mean vector and a log-variance vector, followed by the same clipping and scale computation. The unavailable-state decoders are used only during student training and are removed at inference time.

D Theoretical Notes

This section provides lightweight theoretical notes for the main design choices of CoRe-KD. The goal is not to claim calibrated probabilistic inference

or a formal generalization guarantee. Instead, we use simplified analyses to clarify how complete-view state anchoring, location-scale matching, and NCE can provide structured supervision beyond prediction-level KD or raw reconstruction.

D.1 PoE Fusion as Precision-Weighted State Estimation

The Product-of-Experts fusion used in Eq. (3) can be interpreted as the solution of a precision-weighted state estimation problem. For an available modality set $\mathcal{A}_i^c$, suppose each modality provides a location vector $\boldsymbol{\mu}_{m,i}^c$ and a diagonal precision vector $\boldsymbol{\kappa}_{m,i}^c$. For compactness, define

$$\mathbf{r}_{m,i}^c(\boldsymbol{z}) = \sqrt{\boldsymbol{\kappa}_{m,i}^c} \odot (\boldsymbol{z} - \boldsymbol{\mu}_{m,i}^c), \tag{D.1}$$

where the square root and product are element-wise. Consider the objective

$$\mathcal{J}(\boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{z}\|_2^2 + \frac{1}{2}\sum_{m\in\mathcal{A}_i^c} \|\mathbf{r}_{m,i}^c(\boldsymbol{z})\|_2^2. \tag{D.2}$$

The first term corresponds to a unit Gaussian prior. Setting the gradient to zero gives

$$\boldsymbol{\Lambda}_i^c \odot \boldsymbol{z}_i^\star = \boldsymbol{\eta}_i^c, \tag{D.3}$$

where

$$\begin{aligned} \boldsymbol{\Lambda}_i^c &= \mathbf{1} + \sum_{m\in\mathcal{A}_i^c} \boldsymbol{\kappa}_{m,i}^c, \\ \boldsymbol{\eta}_i^c &= \sum_{m\in\mathcal{A}_i^c} \boldsymbol{\kappa}_{m,i}^c \odot \boldsymbol{\mu}_{m,i}^c. \end{aligned} \tag{D.4}$$

Therefore,

$$\begin{aligned} \boldsymbol{z}_i^\star &= (\boldsymbol{\Lambda}_i^c)^{-1} \odot \boldsymbol{\eta}_i^c, \\ (\boldsymbol{\sigma}_i^c)^2 &= (\boldsymbol{\Lambda}_i^c)^{-1}. \end{aligned} \tag{D.5}$$

This is exactly the fused state location and variance in the PoE fusion rule. This interpretation clarifies that unavailable modalities are omitted from the summation, while modalities assigned higher precision have larger influence on the fused location. The unit prior term acts as a weak reference precision and helps stabilize fusion when only a few modalities are available.

D.2 Location-Scale Matching Geometry

CoRe-KD does not require the Gaussian-inspired states to be calibrated posteriors. We use them as a structured location-scale interface for teacher-reference matching. For two diagonal Gaussian-inspired states

$$q = \mathcal{G}(\boldsymbol{\mu}, \boldsymbol{\sigma}), \qquad q' = \mathcal{G}(\boldsymbol{\mu}', \boldsymbol{\sigma}'),$$

the squared 2-Wasserstein distance has the closed form

$$W_2^2(q, q') = \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2 + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}'\|_2^2. \tag{D.6}$$

Thus, the state distance used in CoRe-KD can be written as

$$\mathcal{D}_G(q, q') = \frac{1}{d} W_2^2(q, q'). \tag{D.7}$$

Compared with matching only $\boldsymbol{\mu}$, this distance also encourages the student to match the teacher-side scale structure.

This state distance is also connected to student-side prediction consistency under a fixed classifier. Let

$$\boldsymbol{\mu}_s = \boldsymbol{\mu}_i^{s,c}, \qquad \boldsymbol{\mu}_t = \boldsymbol{\mu}_i^t.$$

If the student classifier $C^s(\cdot)$ is L-Lipschitz with respect to the fused state location, then

$$\begin{aligned} \|C^s(\boldsymbol{\mu}_s) - C^s(\boldsymbol{\mu}_t)\|_2 &\le L\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\|_2 \\ &\le L\sqrt{d\,\mathcal{D}_G(q_i^{s,c}, q_i^t)}. \end{aligned} \tag{D.8}$$

Therefore, reducing the fused-state distance constrains the logit change induced by the student's fused-state location. Prediction-level anchoring separately aligns the student output with the complete-view teacher logits. This does not by itself guarantee higher classification accuracy, but it provides one explanation for why state anchoring can connect complete-view teacher references with incomplete-view predictions.

D.3 Why Raw Feature Reconstruction Can Be Ambiguous

A common way to handle missing modalities is to reconstruct missing raw or feature-level inputs. However, under squared reconstruction loss, the optimal predictor is the conditional mean of the missing modality given the observed modalities. Let $\mathbf{X}_{\text{obs}}$ denote observed modalities and $\mathbf{X}_{\text{mis}}$ denote missing modalities. The reconstruction objective is

$$\min_g \mathbb{E}\left[\|g(\mathbf{X}_{\text{obs}}) - \mathbf{X}_{\text{mis}}\|_2^2\right]. \tag{D.9}$$

Its Bayes optimal solution is

$$g^\star(\mathbf{X}_{\text{obs}}) = \mathbb{E}[\mathbf{X}_{\text{mis}} \mid \mathbf{X}_{\text{obs}}]. \tag{D.10}$$

When the conditional distribution $p(\mathbf{X}_{\text{mis}} \mid \mathbf{X}_{\text{obs}})$ is multi-modal, this conditional mean may average over several plausible completions. Such an

averaged reconstruction may not correspond to a realistic modality observation and may dilute label-relevant nonverbal evidence.

This analysis does not imply that reconstruction is always ineffective. Rather, it motivates using complete-view teacher states as task-oriented targets when raw missing inputs are not uniquely identifiable. CSA avoids requiring a unique raw completion and instead asks the student to match complete-view fused and modality-specific state references derived from the teacher's task prediction pathway.

D.4 NCE as Target-over-Donor Margin Regularization

NCE constructs target-preserving conflict views by keeping the target language unchanged while replacing nonverbal modalities with a different-label donor. Let y_i be the target label and $y_j \neq y_i$ be the donor label. For a conflict view with replaced modality set b, let $\widetilde{\mathbf{o}}_i^{s,b}$ be the student logits. Define the target-over-donor margin as

$$\Delta_{ij}^{b} = \widetilde{o}_{i,y_i}^{s,b} - \widetilde{o}_{i,y_j}^{s,b}. \tag{D.11}$$

For compactness, define

$$\delta_{ik}^{b} = \widetilde{o}_{i,k}^{s,b} - \widetilde{o}_{i,y_i}^{s,b}. \tag{D.12}$$

The conflict-view cross-entropy is

$$\mathrm{CE}_i^b = \log\left(1 + \sum_{k \neq y_i} \exp(\delta_{ik}^{b})\right). \tag{D.13}$$

Since $y_j \neq y_i$, the donor-label term is included in the summation. Therefore,

$$\mathrm{CE}_i^b \geq \log\left(1 + \exp(-\Delta_{ij}^{b})\right). \tag{D.14}$$

Minimizing the conflict-view CE therefore implicitly encourages a larger target-over-donor margin. This gives a margin-based interpretation of NCE: the model is trained to keep the target label preferred even when nonverbal cues are replaced by donor-label evidence.

This result only explains the optimization effect of NCE under the constructed conflict views. It does not claim that donor swapping exhaustively models all real conversational asynchrony or nonverbal mismatch. In our experiments, NCE is used as a controlled stress regularizer for misleading nonverbal evidence under the evaluated construction.

E Controlled State-Interface Comparison

This section provides controlled comparisons with simpler auxiliary and teacher-guided objectives. The goal is to examine whether the gains of CoRe-KD come from structured complete-view state anchoring rather than from prediction-level KD, generic hidden-state regression, raw feature reconstruction, or regularizer stacking. All variants in this section use the same backbone, input features, missing-modality training views, checkpoint selection rule, and evaluation masks. Unless otherwise specified, the comparison is conducted on IEMOCAP-6, which is the main controlled-analysis dataset.

E.1 Compared Variants

KD only. The student is trained with task CE and prediction-level KD from the frozen complete-view teacher.

KD + Hidden MSE. The student additionally regresses the teacher hidden representation using an MSE loss. This variant tests whether generic representation matching can explain the gains.

KD + Raw Feature Reconstruction. The student additionally reconstructs missing modality features from the available-view representation. This variant tests whether raw feature reconstruction is a sufficient target under the same backbone and training recipe.

KD + Deterministic State. The student matches deterministic state locations without modeling the scale term. This variant tests whether point-vector state regression is sufficient for complete-view guidance.

KD + Gaussian CSA. The student matches both the teacher state location and scale through the Gaussian-inspired state distance. This variant isolates complete-view state anchoring without unavailable-modality-state anchoring or NCE.

CoRe-KD. The full model combines prediction anchoring, Gaussian complete-view state anchoring, unavailable-modality-state anchoring, and NCE.

E.2 Analysis

Table E.1 compares prediction-level KD with stronger teacher-guided objectives under the same IEMOCAP-6 controlled-analysis setting. KD only

Variant	KD	Hidden MSE	Raw Rec.	Det. State	Gaussian CSA	M-State	NCE	Fixed Avg F1	RMFM Avg F1	High-Missing Avg F1
KD only	Yes	No	No	No	No	No	No	56.89 ± 1.26	68.12 ± 0.46	65.52 ± 0.52
KD + Hidden MSE	Yes	Yes	No	No	No	No	No	58.62 ± 0.93	68.41 ± 0.53	65.84 ± 0.86
KD + Raw Feature Rec.	Yes	No	Yes	No	No	No	No	59.24 ± 0.59	69.30 ± 0.58	66.89 ± 0.75
KD + Deterministic State	Yes	No	No	Yes	No	No	No	59.63 ± 0.25	69.43 ± 0.40	67.14 ± 0.99
KD + Gaussian CSA	Yes	No	No	No	Yes	No	No	59.83 ± 0.18	69.68 ± 0.75	67.46 ± 1.36
CoRe-KD	Yes	No	No	No	Yes	Yes	Yes	$\mathbf{62.85 \pm 1.41}$	$\mathbf{71.04 \pm 0.76}$	$\mathbf{69.26 \pm 1.32}$

Table E.1: Controlled comparison with simpler auxiliary and teacher-guided objectives on IEMOCAP-6. Fixed Avg denotes the average weighted F1 over fixed-missing conditions. RMFM Avg denotes the average weighted F1 over random-missing rates. High-Missing Avg denotes the average weighted F1 over high missing rates, i.e., $r \in \{0.5, 0.6, 0.7\}$. Scores are reported as mean $\pm$ standard deviation over five random seeds. M-State denotes unavailable-modality-state anchoring. The full CoRe-KD row is aligned with the final main-setting aggregate.

is already a strong baseline, while hidden-state regression and raw feature reconstruction provide additional gains. State-based variants further improve the three averaged metrics, with Gaussian CSA outperforming deterministic state matching by matching both location and scale. Full CoRe-KD achieves the best results, improving over KD only by 5.96, 2.92, and 3.74 F1 points on Fixed Avg, RMFM Avg, and High-Missing Avg, respectively. It also improves over Gaussian CSA by 3.02, 1.36, and 1.80 points. These results support that the final gains come from structured complete-view state anchoring and complementary robustness regularization, rather than from prediction-level KD or generic auxiliary matching alone.

E.3 Location-Only vs. Location-and-Scale Matching

To further examine the effect of the Gaussian-inspired parameterization, we compare matching only the state location $\boldsymbol{\mu}$ against matching both location and scale $(\boldsymbol{\mu}, \boldsymbol{\sigma})$. Both variants use the same full CoRe-KD setting except for the state-matching distance. The location-only variant replaces the Gaussian-inspired state distance with an MSE loss on $\boldsymbol{\mu}$, while the location-and-scale variant uses the same Gaussian-inspired state distance as the final CoRe-KD model.

Table E.2 isolates the role of the scale term in the Gaussian-inspired state distance. Matching both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ improves over location-only matching by 2.63, 1.39, and 1.72 F1 points on Fixed Avg, RMFM Avg, and High-Missing Avg, respectively. This suggests that the scale term provides useful information beyond point-vector location matching under the same final CoRe-KD setting.

Variant	Fixed Avg F1	RMFM Avg F1	High-Missing Avg F1
Match $\boldsymbol{\mu}$ only	60.22 ± 0.43	69.65 ± 0.40	67.54 ± 0.25
Match $\boldsymbol{\mu}, \boldsymbol{\sigma}$	$\mathbf{62.85 \pm 1.41}$	$\mathbf{71.04 \pm 0.76}$	$\mathbf{69.26 \pm 1.32}$

Table E.2: Effect of matching only the state location versus matching both state location and scale on IEMOCAP-6. All columns report average F1: Fixed Avg over fixed-missing conditions, RMFM Avg over random-missing rates, and High-Missing Avg over $r \in \{0.5, 0.6, 0.7\}$. Scores are reported as mean $\pm$ standard deviation over five random seeds. The $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ row corresponds to the final CoRe-KD state-matching distance and is aligned with the main aggregate results.

Method	A-conflict	V-conflict	AV-conflict	Avg
KD only	61.76	73.32	60.26	65.11
w/o $\mathcal{L}_{\text{NCE}}$	62.78	72.81	61.40	65.66
CoRe-KD	63.69	72.87	62.71	66.42

Table F.1: Controlled target-preserving nonverbal conflict results. All columns report F1.

F Nonverbal Conflict Analysis

F.1 Controlled Conflict Evaluation

We further evaluate CoRe-KD under controlled target-preserving nonverbal conflict views. For each target utterance, the language modality is kept unchanged, while acoustic, visual, or acoustic-visual observations are replaced by different-label donor samples. The target label is preserved during evaluation. This protocol provides a controlled diagnostic for whether the student remains anchored to the target utterance when nonverbal observations introduce misleading donor-label evidence. Unlike the state-space rejection metric in the main analysis, this evaluation reports task-level F1 under conflict views.

As shown in Table F.1, CoRe-KD achieves the best average F1 under the controlled conflict views. Compared with KD only, CoRe-KD improves the

Donor Type	A-conflict	V-conflict	AV-conflict	Rej. Rate
Random different-label donor	63.69	72.87	62.71	95.36
Same-speaker donor	64.17	72.95	62.56	94.38
Same-dialogue donor	63.46	73.34	62.87	82.61
Semantically close donor	63.60	72.87	62.38	85.85

Table F.2: Sensitivity analysis for different donor selection strategies. Conflict columns report F1, and Rej. Rate denotes the fraction of conflict-view states closer to the target anchor than to the donor anchor.

average F1 from 65.11 to 66.42, with clear gains under A-conflict and AV-conflict. Compared with the variant without $\mathcal{L}_{\mathrm{NCE}}$, CoRe-KD further improves the average F1 from 65.66 to 66.42. The V-conflict result is comparable to KD only rather than uniformly better, suggesting that the gain mainly comes from improved robustness to acoustic and audio-visual conflicts. Overall, this task-level evaluation complements the state-space analysis and suggests that nonverbal conflict exposure can help the student preserve target-side evidence when misleading nonverbal cues are introduced.

F.2 Donor Selection Sensitivity

We additionally examine whether the conflict analysis depends on a single donor construction strategy. Specifically, we compare random different-label donors with different-label donors further constrained to come from the same speaker, the same dialogue, or semantically close utterances. This analysis is intended as a sensitivity check rather than as a claim that donor swapping exhaustively models real conversational asynchrony or all forms of nonverbal conflict.

Table F.2 shows that the task-level F1 remains stable across different donor construction strategies. Averaged over the three conflict views, the F1 varies only slightly across random different-label, same-speaker, same-dialogue, and semantically close donors. The rejection rate decreases for same-dialogue and semantically close donors, suggesting that these donors form harder conflict cases, but the corresponding F1 remains close to the random-donor setting. These results suggest that the observed conflict robustness is not merely an artifact of one arbitrary donor choice, while leaving broader real-world nonverbal asynchrony as a more general setting beyond this controlled diagnostic.

G Mechanism Metrics

The metrics in this section are diagnostic measures for examining whether the learned states behave consistently with the intended anchoring behavior. They are not used as independent proof of the method, since they are defined with respect to the same teacher-state interface used during training. We therefore interpret these diagnostics together with the external F1 results and the controlled comparisons in Appendix E.

G.1 Fused-State Drift

We use fused-state drift to measure how far the student fused state deviates from the complete-view teacher state under incomplete observations. For a test condition c, fused-state drift is computed as

$$\mathrm{Drift}(c) = \mathbb{E}_i \left[\mathcal{D}_G \left(q_i^{s,c}, q_i^t \right) \right], \tag{G.1}$$

where $q_i^{s,c}$ is the student fused Gaussian-inspired state under condition c, q_i^t is the complete-view teacher state, and $\mathcal{D}_G(\cdot,\cdot)$ is the state distance defined in the main text. Lower drift indicates closer alignment with the complete-view teacher reference.

G.2 Conflict-View Anchors

For target-preserving nonverbal conflict analysis, we use the complete-view teacher states of the target and donor samples as anchors. Given a target sample i, a different-label donor j, and a replaced nonverbal modality set b, the student produces a conflict-view state $\widetilde{q}_i^{s,b}$. Following the main analysis, we use the state-location distance $d_\mu(q, q') = \|\boldsymbol{\mu}(q) - \boldsymbol{\mu}(q')\|_2$ and define the target-anchor and donor-anchor distances as

$$\begin{aligned} d_{\mathrm{tar}}^{i,b} &= d_\mu \left(\widetilde{q}_i^{s,b}, q_i^t \right), \\ d_{\mathrm{don}}^{i,j,b} &= d_\mu \left(\widetilde{q}_i^{s,b}, q_j^t \right). \end{aligned} \tag{G.2}$$

G.3 Rejection Rate

The rejection indicator measures whether the conflict-view state remains closer to the target anchor than to the donor anchor:

$$r_{i,j}^b = \mathbb{I} \left[d_{\mathrm{tar}}^{i,b} < d_{\mathrm{don}}^{i,j,b} \right]. \tag{G.3}$$

The rejection rate is the average of $r_{i,j}^b$ over valid target–donor pairs and conflict types:

$$\mathrm{Rej} = \mathbb{E}_{i,j,b} \left[r_{i,j}^b \right]. \tag{G.4}$$

A higher rejection rate indicates that the conflict-view state is more often closer to the target teacher anchor than to the donor teacher anchor.

G.4 State and Logit Margins

We also report the target-over-donor state-location margin:

$$m_{i,j}^{b} = d_{\mathrm{don}}^{i,j,b} - d_{\mathrm{tar}}^{i,b}. \tag{G.5}$$

A positive margin means that the conflict-view state is closer to the target teacher anchor than to the donor teacher anchor.

For prediction-level conflict analysis, we use the target-versus-donor logit margin:

$$\Delta_{ij}^{b} = \widetilde{o}_{i,y_i}^{s,b} - \widetilde{o}_{i,y_j}^{s,b}. \tag{G.6}$$

A larger Δ_{ij}^{b} indicates lower donor-label attraction in the conflict view.

H Supplementary IEMOCAP-4 Results

We report supplementary IEMOCAP-4 results under fixed-missing and random-missing protocols. This setting follows the commonly used four-class IEMOCAP protocol and is included to provide additional comparison with prior conversational MER studies. As shown in Table H.1, CoRe-KD achieves the best Acc/F1 in all fixed-missing conditions and all random-missing rates. The gains are especially large when only acoustic or visual observations are available, and also increase under higher random-missing rates. This pattern is consistent with the main IEMOCAP-6 results, suggesting that complete-view state anchoring provides robust supervision across different IEMOCAP label protocols. This table provides the per-condition source results for the IEMOCAP-4 setting, complementing the averaged scores in Table B.1.

I Full Ablation

Table I.1 extends the component ablation results to IEMOCAP-4, MELD-7, and CMU-MOSEI. The CoRe-KD rows are aligned with the corresponding main-table and supplementary-table results. Across these additional settings, the full model remains strongest in most fixed- and random-missing conditions. Removing $\mathcal{L}_{\mathrm{CSA}}$ or $\mathcal{L}_{\mathrm{pred}}$ often leads to large drops, especially in low-information conditions where only nonverbal modalities are available. Removing $\mathcal{L}_{\mathrm{state}}$, $\mathcal{L}_{\mathrm{mstate}}$, or $\mathcal{L}_{\mathrm{NCE}}$ also weakens performance in many cases, suggesting that fused-state alignment, unavailable-modality-state anchoring, and nonverbal conflict regularization provide complementary gains. These extended results provide additional checks that the component-wise trends observed on IEMOCAP-6 are not restricted to a single dataset setting.

J Computational Cost

CoRe-KD introduces additional computation only during training. The complete-view teacher contains 84.8M parameters and is trained once before student training. During student training, teacher references and NCE conflict views increase the training time from 0.14 hours to 0.40 hours on a single NVIDIA RTX A5000 GPU, corresponding to an additional 183.3% training cost. At inference time, the complete-view teacher, unavailable-state decoders, auxiliary training heads, and NCE donor construction are discarded. The deployed model contains 88.0M parameters, where the student-side state heads and PoE fusion module introduce only 8.3M additional parameters, accounting for 9.4% of the deployed model. The inference latency is 30.3 ms per batch, compared with 28.8 ms for the baseline student under the same batch size. Therefore, CoRe-KD introduces no teacher-side computation at deployment time, and its remaining overhead is limited to the lightweight state heads and PoE fusion used by the student.

K Additional t-SNE Visualization under Random Missing

To provide a qualitative view of representation robustness, we further visualize the fused representations on the IEMOCAP-6 test set under random modality feature missing (RMFM) in one seed. The RMFM rate is varied from 0.1 to 0.7, where larger rates indicate more severe sample-wise modality incompleteness. For each RMFM rate, we compare Corr-KD, LNLN, MoMKE, and CoRe-KD under the same visualization protocol. Each point denotes one test utterance and is colored by its ground-truth emotion label. The weighted F1 score of each setting is also reported in the corresponding panel.

As shown in Figure K.1, the baseline methods tend to exhibit increasingly mixed class layouts as the missing rate increases, especially under high RMFM rates such as 0.5, 0.6, and 0.7. In contrast, CoRe-KD maintains a more coherent representation structure across different missing rates, with visually clearer class regions and less severe mixing among several emotion categories. This trend is consistent with the RMFM performance reported in the main results, where CoRe-KD degrades more slowly under increasing modality incompleteness. These qualitative results suggest that complete-view state anchoring helps the student preserve a more stable fused state space under incomplete

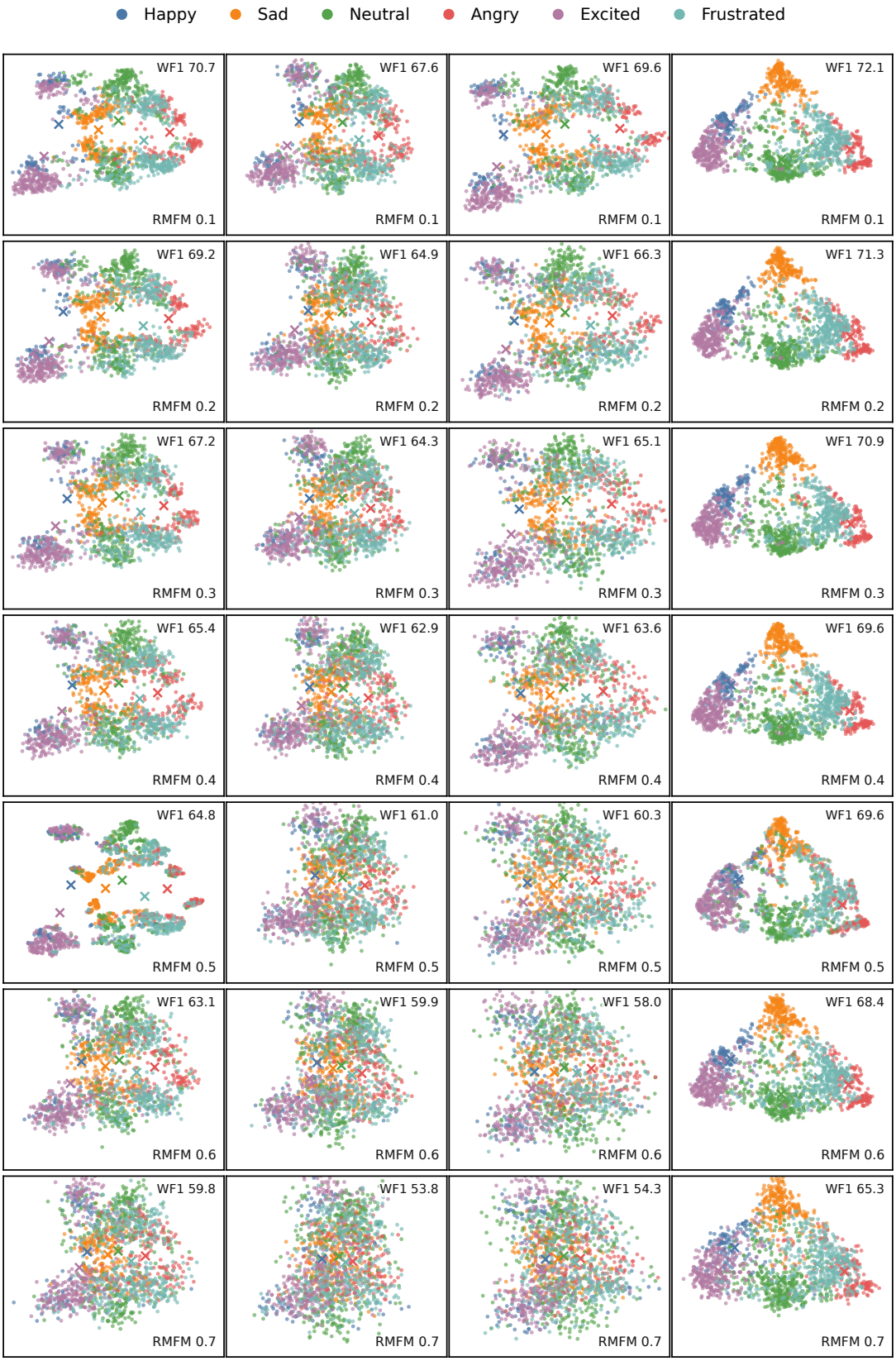


Figure K.1: **t-SNE visualization on IEMOCAP-6 under random modality feature missing.** Rows correspond to RMFM rates from 0.1 to 0.7, and columns correspond to Corr-KD, LNLN, MoMKE, and CoRe-KD, respectively. Each point denotes one test utterance and is colored by its ground-truth emotion label. Each panel reports the corresponding weighted F1 score. The visualization provides a qualitative supplement to the quantitative RMFM results by showing how the fused representation space changes as modality incompleteness becomes more severe.

Testing Condition		$\{a\}$	$\{v\}$	$\{\ell\}$	$\{\ell, a\}$	$\{a, v\}$	$\{\ell, v\}$	$\{\ell, a, v\}$
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Four	IMDer	57.07/55.31	44.12/38.28	78.97/78.87	82.03/82.10	61.40/60.64	80.54/80.39	82.85/82.86
	Corr-KD	68.61/68.20	43.98/35.29	84.87/84.81	86.85/86.91	70.47/69.94	85.80/85.68	87.62/87.65
	MMIN	62.78/61.92	50.87/46.11	78.59/78.57	82.19/82.24	70.07/69.82	80.77/80.65	83.67/83.64
	LNLN	57.64/55.82	49.28/42.90	78.69/78.66	81.92/82.04	66.97/66.58	80.59/80.42	83.68/83.71
	MCULoRA	61.19/59.44	46.75/38.43	80.09/80.12	81.91/82.10	72.47/72.37	82.54/82.28	85.44/85.46
	ComP	52.02/48.71	51.41/43.12	80.12/80.18	83.34/83.44	68.83/68.21	83.31/83.15	85.72/85.67
	CoRe-KD	**76.17/75.83**	**61.43/60.32**	**84.94/84.90**	**88.73/88.67**	**80.68/80.79**	**86.55/86.44**	**89.21/89.03**
	ΔSOTA	↑ 7.56/7.63	↑ 10.02/14.21	↑ 0.07/0.09	↑ 1.88/1.76	↑ 8.21/8.42	↑ 0.75/0.76	↑ 1.59/1.38
Missing Rate		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Dataset	Method	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Four	IMDer	81.79/81.80	80.37/80.41	79.37/79.43	77.68/77.75	75.25/75.33	71.95/72.00	66.31/66.29
	Corr-KD	87.11/87.15	86.24/86.28	85.58/85.61	84.07/84.09	83.04/83.06	81.80/81.77	80.06/79.89
	MMIN	82.68/82.64	81.12/81.11	79.88/79.89	78.64/78.66	76.36/76.39	73.39/73.41	68.47/68.41
	LNLN	82.70/82.70	81.15/81.18	80.09/80.13	78.63/78.69	76.43/76.46	73.48/73.41	68.67/68.46
	MCULoRA	84.35/84.37	83.42/83.44	81.91/81.93	80.73/80.75	78.41/78.41	76.86/76.81	73.62/73.53
	ComP	84.63/84.58	83.69/83.65	82.37/82.33	81.08/81.03	78.80/78.75	76.28/76.25	72.50/72.45
	CoRe-KD	**88.73/88.63**	**88.33/88.22**	**87.12/86.99**	**86.23/86.09**	**86.72/86.63**	**85.02/84.90**	**83.98/83.91**
	ΔSOTA	↑ 1.62/1.48	↑ 2.09/1.94	↑ 1.54/1.38	↑ 2.16/2.00	↑ 3.68/3.57	↑ 3.22/3.13	↑ 3.92/4.02

Table H.1: Supplementary results on IEMOCAP-4 under fixed-missing and random-missing protocols. Best results are bolded and second-best results are underlined.

observations, while the conflict-aware training objective further regularizes the model against unreliable nonverbal evidence. We emphasize that this visualization is used as a qualitative diagnostic rather than independent proof of representation separability.

L Ethics and Limitations

CoRe-KD is evaluated on existing public emotion-recognition datasets and does not introduce new data collection. However, emotion recognition remains a sensitive application area. Emotion labels are subjective and may depend on culture, context, annotator background, and speaker identity. Therefore, strong performance on benchmark splits should not be interpreted as reliable recognition of internal emotional states in real-world deployment.

Robustness under missing modalities can make deployment in low-quality sensing conditions more tempting. This may increase the risk of overclaiming reliability when language, acoustic, or visual signals are noisy, incomplete, or socially ambiguous. In addition, nonverbal expressions vary across speakers, communities, cultures, and disabilities. A model trained to suppress conflicting nonverbal evidence may inadvertently treat some communication styles as unreliable or noisy.

For these reasons, CoRe-KD should not be used for high-stakes profiling, surveillance, hiring, workplace monitoring, education monitoring, or other consequential decision-making applications without strong safeguards, domain validation, and human oversight. The method is intended as a research contribution for robust multimodal learning under controlled benchmark settings.

Testing Condition		$\{a\}$	$\{v\}$	$\{\ell\}$	$\{\ell, a\}$	$\{a, v\}$	$\{\ell, v\}$	$\{\ell, a, v\}$
Dataset	Variant	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Four	w/o $\mathcal{L}_{\text{CSA}}$	69.65/68.81	49.76/45.85	82.21/82.24	86.39/86.44	72.54/71.45	84.38/84.16	87.84/87.72
	w/o $\mathcal{L}_{\text{NCE}}$	68.28/67.91	45.57/44.24	83.49/83.14	85.75/85.60	72.54/72.23	84.30/83.93	85.19/84.93
	w/o $\mathcal{L}_{\text{pred}}$	72.06/72.07	51.05/48.49	83.90/83.79	87.04/87.10	78.10/77.95	84.70/84.49	87.52/87.48
	w/o $\mathcal{L}_{\text{state}}$	74.48/74.42	53.46/51.26	82.53/82.62	87.20/87.28	80.35/80.18	84.78/84.71	88.16/88.15
	w/o $\mathcal{L}_{\text{mstate}}$	75.52/75.18	50.00/46.93	83.74/83.78	87.12/87.15	78.18/77.61	84.70/84.61	88.08/88.02
	CoRe-KD	**76.17/75.83**	**61.43/60.32**	**84.94/84.90**	**88.73/88.67**	**80.68/80.79**	**86.55/86.44**	**89.21/89.03**
MELD Seven	w/o $\mathcal{L}_{\text{CSA}}$	18.62/24.07	26.74/29.82	61.65/63.86	60.34/63.41	19.66/24.87	60.84/63.46	59.08/62.30
	w/o $\mathcal{L}_{\text{NCE}}$	43.33/40.56	29.89/29.77	66.05/66.33	65.63/66.02	37.13/37.15	65.36/66.00	65.52/66.24
	w/o $\mathcal{L}_{\text{pred}}$	22.15/25.50	29.50/32.00	60.73/63.26	60.23/63.08	24.60/29.27	61.00/63.52	59.43/62.55
	w/o $\mathcal{L}_{\text{state}}$	48.12/31.27	48.12/31.27	63.60/58.70	63.37/58.18	48.12/31.27	62.80/57.00	63.03/57.15
	w/o $\mathcal{L}_{\text{mstate}}$	48.12/31.27	48.12/31.27	54.94/42.53	55.33/42.83	48.12/31.27	55.17/42.70	55.21/42.73
	CoRe-KD	**49.77/41.08**	**48.35/32.24**	**68.28/67.25**	**68.47/67.35**	**50.19/38.59**	**68.66/67.97**	**68.85/67.41**
CMU MOSEI	w/o $\mathcal{L}_{\text{CSA}}$	72.04/70.35	58.34/58.41	85.09/85.25	85.97/86.06	71.90/71.79	84.95/85.14	85.94/86.08
	w/o $\mathcal{L}_{\text{NCE}}$	70.72/69.59	66.68/64.33	85.53/85.67	85.58/85.74	70.06/69.82	86.10/86.23	85.61/85.76
	w/o $\mathcal{L}_{\text{pred}}$	71.93/69.70	51.16/48.99	85.42/85.57	85.77/85.90	71.52/71.49	84.78/84.99	85.64/85.80
	w/o $\mathcal{L}_{\text{state}}$	66.81/59.69	56.99/57.17	83.54/83.81	84.89/85.09	71.00/69.02	83.16/83.43	84.76/84.97
	w/o $\mathcal{L}_{\text{mstate}}$	70.36/66.63	53.41/52.44	86.21/86.29	86.90/86.90	72.73/72.32	85.94/86.09	86.60/86.69
	CoRe-KD	**72.95/72.39**	**67.01/67.24**	**86.49/86.49**	**86.98/86.98**	**72.81/72.67**	**86.82/86.85**	**87.12/87.16**
Missing Rate		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Dataset	Variant	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)	Acc(%)/F1(%)
IEMOCAP Four	w/o $\mathcal{L}_{\text{CSA}}$	87.44/87.32	86.72/86.60	86.15/86.04	85.51/85.39	85.02/84.88	82.69/82.48	81.48/81.21
	w/o $\mathcal{L}_{\text{NCE}}$	85.67/85.49	85.10/84.95	83.98/83.80	83.41/83.27	83.17/83.01	82.05/81.85	80.60/80.34
	w/o $\mathcal{L}_{\text{pred}}$	87.04/86.99	86.80/86.74	85.91/85.82	85.19/85.09	84.38/84.28	83.82/83.67	82.13/81.87
	w/o $\mathcal{L}_{\text{state}}$	87.04/87.02	86.96/86.93	85.51/85.48	85.02/84.97	84.38/84.32	82.93/82.87	82.13/81.96
	w/o $\mathcal{L}_{\text{mstate}}$	87.52/87.44	87.20/87.12	86.63/86.56	85.67/85.60	85.75/85.64	84.22/84.08	82.85/82.66
	CoRe-KD	**88.73/88.63**	**88.33/88.22**	**87.12/86.99**	**86.23/86.09**	**86.72/86.63**	**85.02/84.90**	**83.98/83.91**
MELD Seven	w/o $\mathcal{L}_{\text{CSA}}$	56.13/59.80	53.33/57.55	49.31/54.26	45.52/50.93	42.30/48.34	37.82/44.35	31.65/38.29
	w/o $\mathcal{L}_{\text{NCE}}$	63.10/63.72	61.00/61.63	58.28/58.80	54.83/55.31	52.61/52.95	50.15/50.58	47.13/47.33
	w/o $\mathcal{L}_{\text{pred}}$	56.90/60.21	54.52/58.30	50.15/54.37	47.43/52.00	43.41/48.29	39.43/44.35	35.13/40.11
	w/o $\mathcal{L}_{\text{state}}$	61.65/55.32	60.19/53.43	58.35/51.18	56.48/48.47	54.83/45.95	53.60/43.86	51.92/40.62
	w/o $\mathcal{L}_{\text{mstate}}$	54.52/42.00	53.83/41.21	53.14/40.41	52.38/39.46	51.95/38.77	51.19/37.71	50.27/36.14
	CoRe-KD	**66.36/65.56**	**64.18/63.28**	**61.84/60.70**	**60.11/57.61**	**58.81/56.55**	**56.32/52.78**	**54.18/48.42**
CMU MOSEI	w/o $\mathcal{L}_{\text{CSA}}$	85.17/85.30	84.15/84.27	82.91/83.02	82.36/82.41	80.74/80.72	78.92/78.77	77.52/77.07
	w/o $\mathcal{L}_{\text{NCE}}$	84.70/84.82	83.68/83.76	82.80/82.82	82.36/82.24	81.07/80.76	80.13/79.59	77.88/76.72
	w/o $\mathcal{L}_{\text{pred}}$	84.98/85.14	83.76/83.94	82.09/82.27	81.70/81.84	79.99/80.12	79.47/79.54	77.85/77.74
	w/o $\mathcal{L}_{\text{state}}$	83.85/84.05	82.80/83.01	82.17/82.36	81.48/81.60	80.35/80.41	79.25/79.23	77.85/77.61
	w/o $\mathcal{L}_{\text{mstate}}$	85.80/85.87	84.51/84.54	83.19/83.19	83.02/82.90	81.04/80.79	79.53/78.98	77.82/76.71
	CoRe-KD	**86.49/86.40**	**85.06/84.89**	**84.23/84.04**	**83.52/83.21**	**82.00/81.54**	**81.10/80.42**	**78.92/77.77**

Table I.1: Extended ablation table of CoRe-KD on IEMOCAP-4, MELD-7, and CMU-MOSEI under fixed-missing testing conditions and random-missing rates. Ablation rows report the corresponding component-removal results, while the CoRe-KD rows are kept aligned with the main-table and supplementary-table results.