INTERACTION COORDINATE DISTILLATION FOR MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sentiment analysis relies not only on modality-specific evidence but also on how linguistic, acoustic, and visual cues interact. Knowledge distillation can transfer the capabilities of a high-capacity multimodal teacher to a compact student, yet existing approaches primarily define supervision over final predictions, latent representations, or relational structures, leaving predictive interactions across modality combinations largely implicit. We propose **Interaction Coordinate Distillation (ICD)**, which treats teacher and student predictions over the seven non-empty modality subsets as set functions and applies an anchored Möbius decomposition to obtain aligned first-, second-, and third-order interaction coordinates. The invertible transform reorganizes the same teacher subset responses into interaction coordinates, where coordinate-wise matching couples errors across related modality subsets through Möbius contrasts. This allows ICD to change the structure of the distillation constraint without changing the underlying teacher information. ICD jointly optimizes interaction-coordinate matching, task supervision, and full-modality distillation. Experiments on CMU-MOSEI and CMU-MOSI, together with controlled comparisons over raw subset supervision, interaction orders, and random orthogonal reparameterization, show that ICD effectively reconstructs teacher-defined interaction coordinates while achieving strong downstream performance. The additional subset computation required by ICD is absent at inference, introducing no inference-time overhead.

1 INTRODUCTION

Multimodal sentiment analysis (MSA) aims to understand human affect by jointly modeling linguistic, acoustic, and visual signals. Multimodal sentiment depends not only on information from individual modalities but also on how their evidence complements, reinforces, or modifies one another. Prior studies model such dependencies through dynamic fusion, cross-modal attention, and shared–specific representation learning (Zadeh et al., 2018; Tsai et al., 2019; Hazarika et al., 2020), while recent work further emphasizes structured cross-modal dependencies (Meng et al., 2026). Meanwhile, increasingly capable multimodal models incur substantial computational and memory costs, motivating the transfer of their capabilities to compact students (Zheng et al., 2026).

Knowledge distillation (KD) provides a natural mechanism for such transfer. Classical KD aligns teacher and student predictions (Hinton et al., 2015), while subsequent methods extend distillation to intermediate representations and relational structures (Romero et al., 2015; Park et al., 2019). Multimodal approaches further transfer cross-modal representations, sample- and category-level relations, higher-order feature correlations, and pairwise interactions (Li et al., 2023; 2024b; Zhuang et al., 2025; Xu et al., 2026). Prior multimodal distillation predominantly defines such interaction knowledge in latent representation or relational spaces. A different interaction structure is available directly from the teacher's predictive behavior across modality combinations.

Given text (T), audio (A), and vision (V), a model produces responses for the seven non-empty subsets $\{T, A, V, TA, TV, AV, TAV\}$, jointly defining a prediction function over the modality-subset space. Directly distilling these responses provides richer supervision than full-modality KD but treats them as independent targets, even though multimodal responses contain contributions already present in lower-order subsets. For example, TA reflects both unimodal effects and any residual effect emerging from their combination. This motivates a controlled question: *when teacher infor-*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mation is fixed, can reorganizing the same subset responses into interaction-structured coordinates provide a more effective learning constraint than matching them independently?

To investigate this question, we represent the subset prediction function in an interaction coordinate system. Inspired by the Möbius representation of higher-order interactions (Kang et al., 2024), we apply an anchored Möbius decomposition to teacher and student predictions using a shared reference derived from training-set teacher responses. The resulting coordinates characterize modality-specific effects, pairwise residual interactions, and the residual three-way interaction beyond all lower-order terms. For a fixed anchor, the transformation is invertible and preserves exactly the same subset-response information. Its effect arises from the objective: coordinate-wise matching imposes Möbius contrast constraints that couple errors across related modality subsets.

Based on this formulation, we propose **Interaction Coordinate Distillation (ICD)**. During training, teacher and student subset responses are transformed into aligned interaction coordinates and matched coordinate-wise, together with task supervision and full-modality distillation. The student is therefore encouraged to reproduce both the teacher's complete-modality prediction and the residual effects associated with modality combinations. The ICD objective requires subset computation only during training; checkpoint selection, testing, and deployment retain the standard single-TAV student path.

We evaluate ICD on CMU-MOSEI and CMU-MOSI. ICD achieves the best Acc-2, weighted F1, and Pearson correlation among the compared methods on CMU-MOSEI, while on CMU-MOSI it achieves the highest Pearson correlation together with strong performance across the remaining metrics. Controlled comparisons with raw subset supervision, progressively higher interaction orders, and a random orthogonal transformation examine the roles of additional subset information, interaction order, and arbitrary orthogonal reparameterization. We further analyze teacher–student interaction fidelity and deployment efficiency.

The main contributions are:

- We formulate multimodal distillation from a *prediction-function interaction* perspective, treating teacher responses over modality subsets as a unified set function and studying how their supervision structure affects student learning.
- We propose **Interaction Coordinate Distillation (ICD)**, which applies an anchored Möbius decomposition to obtain aligned first-, second-, and third-order interaction coordinates. The transformation induces structured coupling among subset-response errors through aligned interaction coordinates, with no additional inference-time cost.
- We conduct systematic experiments on CMU-MOSEI and CMU-MOSI with controlled comparisons over raw subset supervision, interaction orders, and random orthogonal reparameterization, together with interaction-fidelity and deployment-efficiency analyses.

2 Related Work

2.1 Multimodal Sentiment Analysis and Knowledge Distillation

Multimodal sentiment analysis (MSA) integrates textual, acoustic, and visual evidence through mechanisms such as cross-modal attention, shared–private representation learning, and structured multimodal representation spaces (Tsai et al., 2019; Hazarika et al., 2020; Meng et al., 2026). Another line of work considers incomplete or uncertain modalities. MissModal promotes consistent representations across different modality configurations, while UMDF unifies heterogeneous missing-modality states through self-distillation and multi-grained cross-modal interaction (Lin & Hu, 2023; Li et al., 2024a). These studies demonstrate that representations and predictions under different modality combinations contain useful structure beyond the complete multimodal input.

Knowledge distillation transfers multimodal knowledge from high-capacity teachers to compact students. DMD performs adaptive graph-based cross-modal distillation, CorrKD transfers cross-sample, cross-category, and response-level correlations, MCDN adapts teacher knowledge to missing-modality states, and CMAD combines feature alignment with high-order inter-instance correlation distillation (Li et al., 2023; 2024b; Hu et al., 2025; Zhuang et al., 2025). Recent incomplete-modality methods develop structured and multilevel transfer: HCKD aligns structural, semantic, and

decision knowledge, APRD transfers sentiment prototypes and inter-sample relations, and HyPLe-MKD combines prompt-based modality completion with feature- and response-level distillation (Xi et al., 2026; Sun et al., 2026; Zhai et al., 2026b). Other approaches consider uncertain missingness and representation alignment (Wang & Wang, 2026; Luo et al., 2025), pairwise cross-modal correlations (Xu et al., 2026), and inter-sample relational knowledge (Ye et al., 2026; Zhai et al., 2026a), while Light-MER studies large-to-small distillation for multimodal emotion recognition (Zheng et al., 2026). Beyond MSA, CCSD and HMDMV exploit supervision across multiple modality or view combinations (Xie et al., 2026; Yang et al., 2026a). In contrast, ICD operates directly on the teacher's prediction function over modality subsets and reorganizes the same subset responses into interaction-structured distillation targets rather than defining interaction knowledge in a learned feature or relational space.

2.2 Multimodal Interaction Modeling

Multimodal interactions have been characterized from representation, prediction, information-theoretic, and game-theoretic perspectives. TFN explicitly parameterizes unimodal, bimodal, and trimodal feature interactions (Zadeh et al., 2017), while EMAP separates additive contributions from non-additive cross-modal dependencies at the prediction level (Hessel & Lee, 2020). Information-theoretic approaches characterize redundancy, uniqueness, and synergy among modalities (Liang et al., 2023; Wörtwein et al., 2024); LSMI estimates sample-wise information components for targeted distillation, whereas DMIL learns from sample-specific information-theoretic decomposition (Yang et al., 2025b; 2026b). Game-theoretic methods treat modality subsets as coalitions: SHAPE quantifies modality contributions and cooperation, while InterSHAP characterizes pairwise interactions between modalities (Hu et al., 2022; Wenderoth et al., 2025). For general set functions, the Möbius transform provides a complete subset-indexed representation whose higher-order coefficients capture residual effects beyond lower-order subsets and connect naturally to Shapley-style importance measures (Kang et al., 2024). Whereas prior interaction modeling primarily uses such structure for representation learning, interpretation, or attribution, ICD uses it as a distillation coordinate system: teacher and student prediction functions are expressed in the same interaction basis, and the resulting coordinates directly define the training constraints.

3 Method

Given a multimodal sentiment sample $\mathbf{x}_i = (x_i^{\mathrm{T}}, x_i^{\mathrm{A}}, x_i^{\mathrm{V}})$, where T, A, and V denote text, audio, and video, respectively, each sample has a continuous sentiment label $y_i \in [-3, 3]$. We round y_i to the nearest integer, with half-integer ties rounded away from zero, to obtain $\tilde{c}_i \in \{-3, \dots, 3\}$, and use $c_i = \tilde{c}_i + 3 \in \{0, \dots, 6\}$ as the seven-class target. We denote the full modality set by $\mathcal{M} = \{\mathrm{T}, \mathrm{A}, \mathrm{V}\}$ and consider all non-empty subsets,

$$\mathcal{P}^+(\mathcal{M}) = \{\mathrm{T}, \mathrm{A}, \mathrm{V}, \mathrm{TA}, \mathrm{TV}, \mathrm{AV}, \mathrm{TAV}\}.$$

As illustrated in Figure 1, Interaction Coordinate Distillation (ICD) first obtains teacher and student responses over all non-empty modality subsets, then applies a shared anchored Möbius decomposition to construct aligned first-, second-, and third-order interaction coordinates. The resulting teacher coordinates supervise the corresponding student coordinates together with task and full-modality distillation, while using the same seven teacher subset responses as direct subset distillation.

3.1 Subset Response Construction

For each non-empty modality subset $S \in \mathcal{P}^+(\mathcal{M})$, we construct teacher and student sentiment responses $r_i^{\mathcal{T}}(S)$ and $r_i^{\mathcal{S}}(S)$.

Frozen teacher. The teacher consists of a frozen large multimodal model followed by $K = 3$ frozen sentiment probes. For subset S, the teacher receives only the modalities in S, and its last valid input-token state is passed through a LayerNorm–Linear$(2048, 512)$–GELU–Dropout(0.1) probe followed by separate seven-class and scalar-regression heads, with the scalar output parameterized as $3\tanh(\mathrm{Linear}(512, 1))$. The probes are trained only on the official training split with

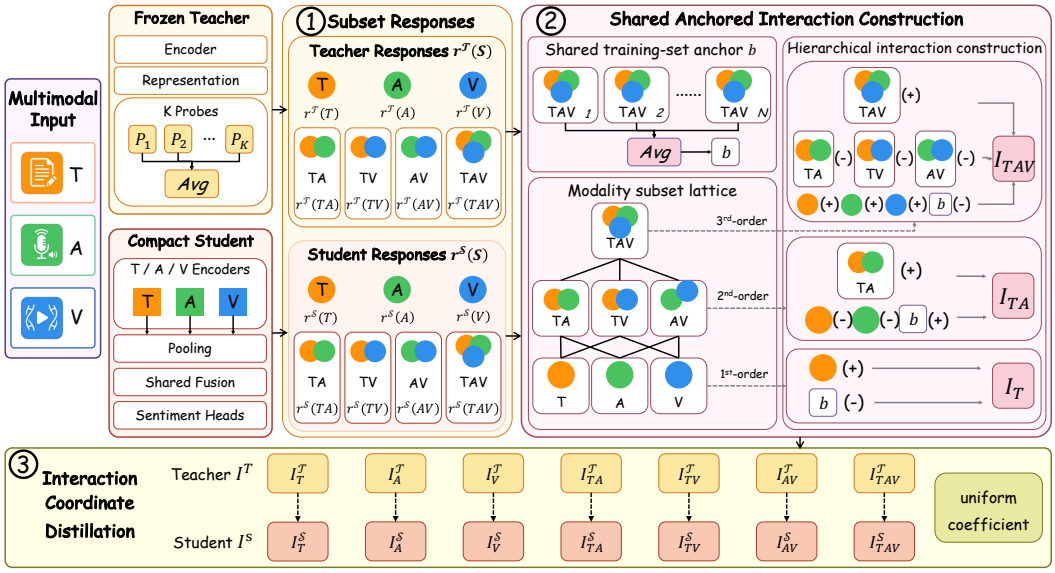


Figure 1: Overview of Interaction Coordinate Distillation (ICD). Teacher and student responses over the same seven non-empty modality subsets are mapped into aligned first-, second-, and third-order Möbius interaction coordinates. ICD matches the corresponding coordinates during training, while deployment retains only the standard full-modality student path.

seeds 2026/2027/2028 and selected by validation joint loss; their seven-class logits are temperature-calibrated on validation NLL. Their scalar responses are averaged as $r_i^{\mathcal{T}}(S) = K^{-1} \sum_{j=1}^{K} r_{i,j}^{\mathcal{T}}(S)$.

Compact student. The student contains modality-specific text, audio, and video encoders, query-based pooling modules, a shared multimodal fusion network, and sentiment heads. Each encoder produces a sequence representation compressed into fixed-length modality tokens through learned-query pooling. For modality $m \in \mathcal{M}$, the fusion network receives its encoded representation when $m \in S$ and learnable missing-modality tokens otherwise, producing $r_i^{\mathcal{S}}(S)$ and logits $\mathbf{z}_i^{\mathcal{S}}(S)$. Each modality is encoded once per sample and reused across all seven subset configurations.

3.2 Shared Anchored Interaction Construction

Given a subset-response function $r(S)$, we construct hierarchical interaction coordinates over the modality-subset lattice using an anchored Möbius decomposition.

Shared training-set anchor. We define the empty-set reference from the aggregated full-modality teacher responses:

$$b = \frac{1}{N_{\text{tr}}} \sum_{i=1}^{N_{\text{tr}}} r_i^{\mathcal{T}}(\text{TAV}), \tag{1}$$

where N_{tr} is the number of training samples. The anchor remains fixed and is shared by the teacher and student.

For a subset-response function $r(\cdot)$, let $\widetilde{r}(\emptyset) = b$ and $\widetilde{r}(U) = r(U)$ for $U \neq \emptyset$. The anchored Möbius operator is

$$\mathcal{A}_b[r](S) = \sum_{U \subseteq S} (-1)^{|S|-|U|} \widetilde{r}(U), \qquad S \in \mathcal{P}^+(\mathcal{M}). \tag{2}$$

The transform yields three first-order coordinates $I_{\text{T}}, I_{\text{A}}, I_{\text{V}}$, three second-order coordinates $I_{\text{TA}}, I_{\text{TV}}, I_{\text{AV}}$, and one third-order coordinate I_{TAV}. For example,

$$I_{\mathrm{T}} = r(\mathrm{T}) - b, \qquad I_{\mathrm{TA}} = r(\mathrm{TA}) - r(\mathrm{T}) - r(\mathrm{A}) + b.$$

We use "interaction" to denote predictive residuals of the modality-subset response function.

Applying the same operator to teacher and student responses gives $I_i^{\mathcal{T}}(S) = \mathcal{A}_b[r_i^{\mathcal{T}}](S)$ and $I_i^{\mathcal{S}}(S) = \mathcal{A}_b[r_i^{\mathcal{S}}](S)$. For fixed b, the transform is invertible and therefore preserves the information in the original subset responses.

Because the same anchor is used for both models, it cancels from their coordinate discrepancy. Defining $\Delta r_i(S) = r_i^{\mathcal{S}}(S) - r_i^{\mathcal{T}}(S)$ gives

$$I_i^{\mathcal{S}}(S) - I_i^{\mathcal{T}}(S) = \sum_{\emptyset \neq U \subseteq S} (-1)^{|S|-|U|} \Delta r_i(U). \tag{3}$$

Thus, the anchor fixes the absolute coordinate reference but cancels from the ICD discrepancy and does not affect its optimization.

3.3 Interaction Coordinate Distillation

The teacher interaction coordinates provide targets for the corresponding student coordinates. We use the same loss coefficient for all seven coordinates:

$$\mathcal{L}_{\mathrm{ICD}} = \mathbb{E}_i \left[\frac{1}{7} \sum_{S \in \mathcal{P}^+(\mathcal{M})} \ell_{\mathrm{Huber}} \left(I_i^{\mathcal{S}}(S), I_i^{\mathcal{T}}(S) \right) \right], \tag{4}$$

where ℓ_{Huber} denotes the Smooth-L1 loss with $\beta = 1$.

Since the transform preserves the subset-response information, the effect of ICD comes from applying the loss in interaction coordinates, which changes how subset-response errors are constrained. Let $\Delta \mathbf{r}_i \in \mathbb{R}^7$ collect the seven response discrepancies and $\mathbf{M} \in \mathbb{R}^{7\times 7}$ denote the fixed Möbius contrast matrix. Then

$$\Delta \mathbf{I}_i = \mathbf{M} \Delta \mathbf{r}_i. \tag{5}$$

Subset-7 KD applies a coordinate-wise loss directly to $\Delta \mathbf{r}_i$, whereas ICD applies it after the structured transformation $\mathbf{M}$. For example,

$$I_{i,\mathrm{TA}}^{\mathcal{S}} - I_{i,\mathrm{TA}}^{\mathcal{T}} = \Delta r_i(\mathrm{TA}) - \Delta r_i(\mathrm{T}) - \Delta r_i(\mathrm{A}),$$

so a pairwise interaction constraint jointly couples three related subset responses, with higher-order coordinates coupling larger subsets analogously.

Training objective. We optimize ICD together with task supervision and full-modality distillation. The task objective combines regression and seven-class classification on the student's full TAV path, while full-modality distillation uses the seed-2026 probe, fixed in advance as the reference teacher $\star$:

$$\begin{aligned} \mathcal{L}_{\mathrm{task}} &= \mathbb{E}_i \left[\ell_{\mathrm{Huber}} \big(r_i^{\mathcal{S}}(\mathrm{TAV}), y_i \big) + \alpha_{\mathrm{ce}} \ell_{\mathrm{CE}} \big(\mathbf{z}_i^{\mathcal{S}}(\mathrm{TAV}), c_i \big) \right], \\ \mathcal{L}_{\mathrm{full}} &= \mathbb{E}_i \left[\ell_{\mathrm{Huber}} \big(r_i^{\mathcal{S}}(\mathrm{TAV}), r_{i,\star}^{\mathcal{T}}(\mathrm{TAV}) \big) + \tau^2 D_{\mathrm{KL}} \big(p_{i,\tau}^{\mathcal{T}} \| p_{i,\tau}^{\mathcal{S}} \big) \right]. \end{aligned} \tag{6}$$

Here, α_{ce} balances regression and classification, and τ is the distillation temperature. We define $p_{i,\tau}^{\mathcal{S}} = \mathrm{softmax}(\mathbf{z}_i^{\mathcal{S}}(\mathrm{TAV})/\tau)$ and $p_{i,\tau}^{\mathcal{T}} = \mathrm{softmax}(\mathbf{z}_{i,\star}^{\mathcal{T}}(\mathrm{TAV})/(\tau_{\mathrm{cal}}\tau))$, where τ_{cal} is the validation-fitted calibration temperature of the fixed reference probe.

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{full}} \mathcal{L}_{\mathrm{full}} + \lambda_{\mathrm{ICD}} \mathcal{L}_{\mathrm{ICD}}, \tag{7}$$

where λ_{full} and λ_{ICD} control the two distillation terms. The full-modality term maintains direct teacher–student matching on the complete input, while $\mathcal{L}_{\mathrm{ICD}}$ imposes complementary constraints on modality-specific and residual interaction coordinates.

During training, each student modality encoder is executed once and reused across all seven subsets, with additional computation confined to the fusion network and prediction heads. Checkpoint selection, testing, and deployment use only the full TAV student path; outside training, additional subset passes are used only for post-hoc interaction analysis.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We use the official train/validation/test splits of CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018). CMU-MOSI contains 1,284/229/686 utterances and CMU-MOSEI contains 16,326/1,871/4,659. Both provide text, audio, and visual modalities with sentiment scores in $[-3, 3]$. We report binary accuracy (Acc-2), weighted F1, mean absolute error (MAE), Pearson correlation, and seven-class accuracy (Acc-7). Following the common non-zero protocol, zero-score samples are excluded from Acc-2 and weighted F1.

Teacher, student, and baselines. The teacher is frozen Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) with three frozen sentiment probes. The student uses Qwen3-0.6B-Base (Yang et al., 2025a), WavLM-Base-Plus (Chen et al., 2022), and VideoMAE-Base (Tong et al., 2022) as text, audio, and visual encoders. It contains 819.31M parameters, of which 42.65M (5.2%) are trainable. Adapted Student uses task supervision only; Full KD performs complete-TAV response distillation (Hinton et al., 2015); Projector Feature KD aligns heterogeneous features through a learned projector (Miles & Mikolajczyk, 2024); EA-KD adaptively modulates distillation according to predictive states (Su et al., 2025); and CMAD-style CAFD adapts the Correlation-Aware Feature Distillation component of CMAD (Zhuang et al., 2025) while preserving the fixed student architecture.

For controlled comparisons, Subset-7 KD matches the same seven teacher responses used by ICD, directly controlling for access to additional subset supervision. First-order supervises only unimodal coordinates, while First+Second-order additionally includes pairwise coordinates, allowing us to examine the effect of interaction order. Random Orthogonal applies a fixed random orthogonal transform to the same subset responses before coordinate-wise matching. Since Smooth-L1 is not rotation invariant, this control measures the effect of an arbitrary orthogonal reparameterization under the same coordinate-wise matching loss. All controlled variants use the same student architecture, preprocessing, and optimization protocol, differing only in the structure of the distillation supervision.

Implementation details. All fixed-student comparisons use identical preprocessing, architecture, and adaptation settings. We use AdamW with batch size 8, learning rate 1×10^{-4}, and weight decay 0.01. Training runs for up to 20 epochs with patience 7, and checkpoints are selected by validation MAE. We use $\alpha_{\text{ce}} = 0.5$, $\tau = 2$, and $\lambda_{\text{full}} = \lambda_{\text{ICD}} = 1$, with a uniform loss coefficient across interaction coordinates and Smooth-L1 ($\beta = 1$). Unless otherwise specified, main comparisons report the common seed-13 protocol, while controlled ablations report mean $\pm$ standard deviation over three seeds. Full KD uses the fixed seed-2026 probe for its TAV target, whereas ICD constructs its seven subset-interaction targets from the mean responses of all three probes.

The three teacher probes are trained independently with AdamW (learning rate 10^{-3}, weight decay 10^{-4}, batch size 128) for up to 200 epochs with patience 25, using weighted Smooth-L1 regression plus 0.5 seven-class cross-entropy, with window weights normalized within each parent–subset group; each minimum-validation-joint-loss checkpoint is subsequently calibrated by scalar temperature scaling on validation NLL. During probe training, T, A, V, TA, TV, and AV are sampled with probability 0.1 each, while TAV is sampled with probability 0.4.

4.2 MAIN RESULTS

Table 1 summarizes the main comparisons on CMU-MOSEI and CMU-MOSI.

CMU-MOSEI. ICD achieves the highest Acc-2, weighted F1, and Pearson correlation, reaching 0.8819, 0.8810, and 0.8222. Its Acc-2 exceeds Full KD and EA-KD by 1.32 and 1.18 percentage points. Projector Feature KD obtains the best MAE (0.4772) and Acc-7 (0.5776), while ICD reaches a nearly identical MAE of 0.4774.

CMU-MOSI. CMAD-style CAFD achieves the best Acc-2, weighted F1, and MAE, whereas ICD obtains the highest Pearson correlation (0.8087) and second-lowest MAE (0.7225).

Table 1: Main results on CMU-MOSEI and CMU-MOSI.

Method	Acc-2↑	F1↑	MAE↓	Pearson↑	Acc-7↑
CMU-MOSEI					
Adapted Student	0.8564	0.8571	0.5284	0.7830	0.5415
Full KD	0.8687	0.8687	0.4816	0.8126	0.5752
Projector Feature KD	0.8671	0.8674	**0.4772**	0.8194	**0.5776**
EA-KD	0.8701	0.8708	0.4811	0.8123	0.5744
CMAD-style CAFD	0.8657	0.8660	0.4845	0.8139	0.5752
ICD (Ours)	**0.8819**	**0.8810**	0.4774	**0.8222**	0.5701
CMU-MOSI					
Adapted Student	0.8247	0.8251	0.7972	0.7514	0.4315
Full KD	0.8460	0.8464	0.7285	0.7961	**0.4548**
Projector Feature KD	0.8415	0.8424	0.7272	0.7884	0.4329
EA-KD	0.8552	0.8555	0.7518	0.7909	0.4155
CMAD-style CAFD	**0.8598**	**0.8602**	**0.7154**	0.7919	0.4519
ICD (Ours)	0.8537	0.8545	0.7225	**0.8087**	0.4417

Table 2: Three-seed controlled ablations on CMU-MOSEI and CMU-MOSI.

	CMU-MOSEI		CMU-MOSI	
Method	Acc-2↑	F1↑	Acc-2↑	F1↑
Full KD	0.8728 ± 0.0061	0.8728 ± 0.0057	0.8399 ± 0.0070	0.8401 ± 0.0071
Subset-7 KD	0.8761 ± 0.0033	0.8752 ± 0.0032	0.8450 ± 0.0162	0.8458 ± 0.0159
First-order	0.8683 ± 0.0056	0.8687 ± 0.0047	0.8425 ± 0.0150	0.8427 ± 0.0148
First+Second-order	0.8736 ± 0.0010	0.8739 ± 0.0006	0.8420 ± 0.0148	0.8425 ± 0.0147
Random Orthogonal	0.8762 ± 0.0077	0.8762 ± 0.0075	0.8471 ± 0.0089	0.8473 ± 0.0087
ICD (Ours)	$\mathbf{0.8778 \pm 0.0044}$	$\mathbf{0.8773 \pm 0.0042}$	$\mathbf{0.8486 \pm 0.0075}$	$\mathbf{0.8493 \pm 0.0076}$

Relative to the task-only Adapted Student, ICD improves Acc-2 by 2.55/2.90 percentage points and Pearson correlation by 0.0392/0.0573 on CMU-MOSEI/CMU-MOSI, respectively, while yielding relative MAE reductions of 9.7%/9.4%.

4.3 Controlled Ablations

We conduct controlled experiments to distinguish three factors that may contribute to the behavior of ICD: access to additional subset supervision, the inclusion of different interaction orders, and arbitrary orthogonal reparameterization. All variants use the same student architecture, preprocessing, and optimization protocol, and Table 2 reports mean $\pm$ standard deviation over three seeds.

Subset supervision. Subset-7 KD and ICD use exactly the same seven teacher subset responses, differing in whether the matching loss is applied directly to the raw responses or to their interaction coordinates. On CMU-MOSEI, Subset-7 reaches 0.8761 ± 0.0033 Acc-2 and 0.8752 ± 0.0032 F1, while ICD reaches 0.8778 ± 0.0044 and 0.8773 ± 0.0042, respectively. The same pattern appears on CMU-MOSI, where ICD improves the corresponding means from 0.8450/0.8458 to 0.8486/0.8493. Under the three-seed protocol, ICD attains higher mean Acc-2 and F1 than direct subset-response matching on both datasets.

Interaction order. First-order supervision yields lower mean Acc-2 and F1 than full ICD on both datasets. On CMU-MOSEI, adding the three pairwise coordinates recovers a substantial portion of the gap, increasing Acc-2 from 0.8683 to 0.8736 and F1 from 0.8687 to 0.8739, although full ICD remains higher in both metrics. On CMU-MOSI, full ICD again achieves the highest mean Acc-2 and F1 among the controlled interaction-coordinate variants.

Coordinate reparameterization. Random Orthogonal reaches 0.8762 ± 0.0077 Acc-2 on CMU-MOSEI and 0.8471 ± 0.0089 on CMU-MOSI, while ICD obtains the highest mean Acc-2 and F1

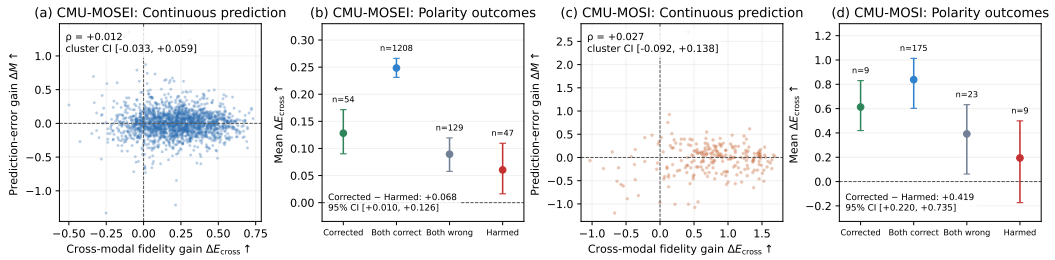


Figure 3: Cross-modal fidelity gains and prediction outcomes on the validation sets. Intervals denote 95% cluster-bootstrap CIs over source videos.

on both datasets. Section 4.4 further compares their reconstruction of teacher-defined Möbius interactions.

4.4 Interaction Analysis

Interaction reconstruction by order. Figure 2 compares teacher–student reconstruction errors for first-, second-, and third-order coordinates on the official validation sets. All methods are evaluated post hoc in the same anchored Möbius space. Overall reconstruction MAE averages the absolute teacher–student errors over all seven interaction coordinates and validation samples.

On CMU-MOSEI, First+Second-order achieves the lowest overall reconstruction MAE (0.2653), while ICD is close at 0.2669 and obtains the lowest third-order error (0.2398). On CMU-MOSI, ICD achieves the lowest overall reconstruction MAE (0.5384). Compared with Random Orthogonal, ICD yields lower Möbius reconstruction error on both datasets, indicating closer alignment with the teacher-defined interaction coordinates.

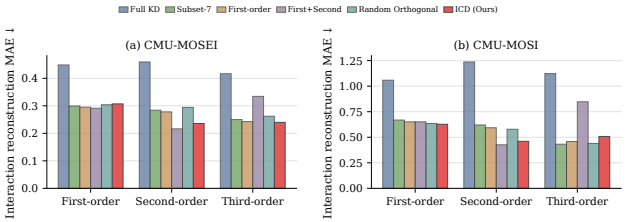


Figure 2: Interaction reconstruction MAE by order on CMU-MOSEI and CMU-MOSI. Lower is better.

Interaction fidelity and prediction outcomes. For each sample i and coordinate S, we define the teacher–student reconstruction error as $e_i(S) = |I_i^{\mathcal{S}}(S) - I_i^{\mathcal{T}}(S)|$. We then define $E_{\text{cross},i} = (e_i(\text{TA}) + e_i(\text{TV}) + e_i(\text{AV}) + e_i(\text{TAV}))/4$ and $\Delta E_{\text{cross},i} = E^{\text{Full}}_{\text{cross},i} - E^{\text{ICD}}_{\text{cross},i}$. We similarly define $\Delta M_i = |y_i - \hat{y}_i^{\text{Full}}| - |y_i - \hat{y}_i^{\text{ICD}}|$, with positive values favoring ICD.

Figure 3(a,c) reports Spearman's ρ values of 0.012 on CMU-MOSEI and 0.027 on CMU-MOSI for the sample-level association between fidelity gain and continuous error improvement, with both bootstrap intervals including zero.

For polarity outcomes, we call samples corrected by ICD but not Full KD *Corrected*, and the reverse cases *Harmed*. The mean fidelity gain of Corrected samples exceeds that of Harmed samples by 0.068 on CMU-MOSEI (95% CI $[0.010, 0.126]$) and 0.419 on CMU-MOSI (95% CI $[0.220, 0.735]$).

Qualitative analysis. Both examples are selected solely by teacher-side interaction statistics. The first is ranked by mixed-sign first-order strength $s_{\text{opp}} = \min(\max_m [I_m]_+, \max_m [-I_m]_+)$ and ranks 6th among 696 eligible utterances; the second is selected from the pre-specified top-5% by $|I_{\text{TA}}|$ and ranks 66th among all 1,871 validation utterances.

The first example contains conflicting first-order effects: positive text and audio contributions but a negative visual contribution. ICD obtains reconstruction MAE 0.614 and predicts $+0.633$ for ground truth $+0.667$ (error 0.034), whereas First-order achieves lower reconstruction MAE (0.525)

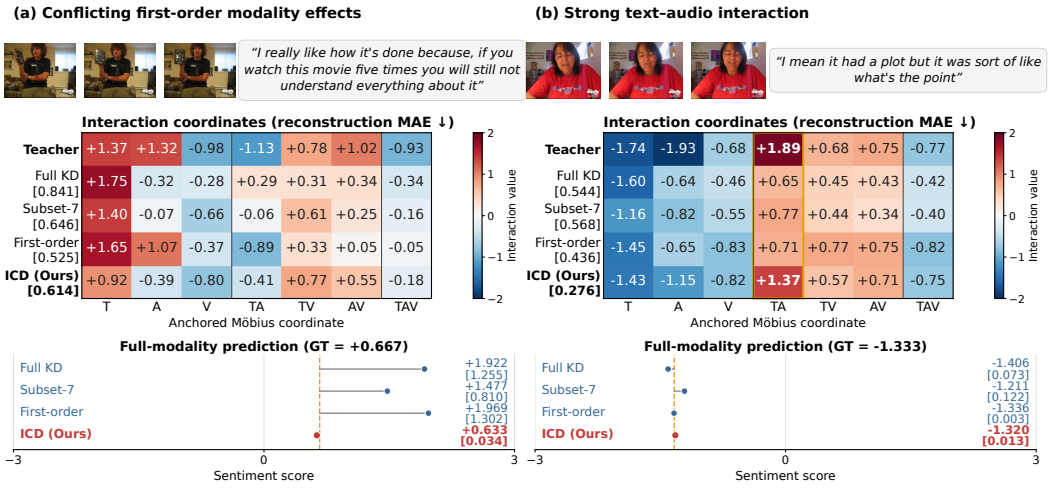


Figure 4: Qualitative interaction reconstruction on two teacher-selected CMU-MOSEI examples. Brackets denote reconstruction MAE and absolute prediction error.

Table 3: Deployment efficiency on a single NVIDIA RTX 6000D with batch size 1.

Model	Params↓	Peak Mem.↓ (GiB)	p50/p95↓ (ms)	E2E↑ (utt/s)
Teacher + Probe	31.720B	67.526	435 / 2104	0.735
Full KD	819.309M	2.091	94 / 127	1.630
ICD (Ours)	819.309M	2.091	95 / 126	1.630

but much larger prediction error (1.302). The second example contains a large teacher text–audio interaction coordinate ($I_{\mathrm{TA}} = +1.89$); ICD reconstructs it as $+1.37$ and obtains the lowest overall reconstruction MAE (0.276).

Deployment efficiency. We measure teacher and student inference sequentially on the same NVIDIA RTX 6000D with batch size 1 over 90 distinct official-validation utterances after three disjoint warm-up samples, with CUDA synchronization before and after each model forward.

Model latency includes CPU-to-GPU transfer and forward computation, while end-to-end throughput additionally includes input reading, video decoding and sampling, tokenization, and multimodal preprocessing. The deployed student reduces parameters by $38.7\times$ and peak allocated memory by $32.3\times$ relative to the teacher while increasing throughput by $2.2\times$; Full KD and ICD have essentially identical inference costs, with no additional deployment-time branch or parameters for ICD. Teacher supervision is generated offline and reused across methods and seeds: on CMU-MOSI, preparing all frozen targets required 89.28 min and approximately 2.96 GPU-hours, while student training required 0.807 GPU-hours for Full KD and 0.806 GPU-hours for ICD, with no teacher forward pass during student training. After offline target preparation, ICD and Full KD therefore have essentially identical student-training cost, differing by less than 0.2%.

5 Conclusion

We introduced Interaction Coordinate Distillation (ICD), which reorganizes the teacher's prediction function over modality subsets into first-, second-, and third-order Möbius interaction coordinates for structured distillation. Experiments on CMU-MOSEI and CMU-MOSI show strong downstream performance with no additional inference-time cost. Controlled comparisons show higher mean performance than direct subset-response matching, while interaction analyses show the lowest third-order reconstruction error on CMU-MOSEI and the lowest overall reconstruction error on CMU-MOSI.

AI USE STATEMENT

Large language models were used for literature discovery, manuscript drafting and polishing, and feedback on research framing and experimental design. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the paper.

ETHICS STATEMENT

This work uses established multimodal sentiment benchmarks and collects no new human-subject data. Sentiment models may inherit biases from their data and pretrained models and should be evaluated carefully in sensitive applications.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 specify the main model, training, evaluation, and checkpoint-selection details. Additional ablation, interaction-analysis, and deployment protocols are described in the corresponding experimental sections.

REFERENCES

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131. Association for Computing Machinery, 2020. doi: 10.1145/3394171.3413678.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.62.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Pengbo Hu, Xingyu Li, and Yi Zhou. SHAPE: An unified approach to evaluate the contribution and cooperation of individual modalities. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3064–3070, 2022. doi: 10.24963/ijcai.2022/425.

Zhangfeng Hu, Wenming Zheng, Mengting Wei, Mengxin Shi, and Yuan Zong. Missing customized distillation network for incomplete multimodal sentiment analysis. In *Pattern Recognition. ICPR 2024*, volume 15308 of *Lecture Notes in Computer Science*, pp. 49–64, Cham, 2025. Springer Nature Switzerland. doi: 10.1007/978-3-031-78186-5_4.

Justin Singh Kang, Yigit Efe Erginbas, Landon Butler, Ramtin Pedarsani, and Kannan Ramchandran. Learning to understand: Identifying interactions via the Möbius transform. In *Advances in Neural Information Processing Systems*, volume 37, pp. 46160–46202. Curran Associates, Inc., 2024. doi: 10.52202/079017-1467.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. 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, 2024a. doi: 10.1609/aaai.v38i9.28871.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12458–12468, 2024b. doi: 10.1109/CVPR52733.2024.01184.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6631–6640, 2023. doi: 10.1109/CVPR52729.2023.00641.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27351–27393. Curran Associates, Inc., 2023. doi: 10.52202/075280-1192.

Ronghao Lin and Haifeng Hu. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023. doi: 10.1162/tacl_a_00628.

Yanqian Luo, Shijin Wang, Zhongxing Xu, Yulong Li, Feilong Tang, and Jionglong Su. Confidence-aware self-distillation for multimodal sentiment analysis with incomplete modalities. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11208961.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8791–8800, 2026.

Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4233–4241, 2024. doi: 10.1609/aaai.v38i5.28219.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Chi-Ping Su, Ching-Hsun Tseng, Bin Pu, Lei Zhao, Jiewen Yang, Zhuangzhuang Chen, and Shin-Jye Lee. EA-KD: Entropy-based adaptive knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 731–740, 2025. doi: 10.1109/ICCV51701.2025.00076.

Xiyang Sun, Linkai Luo, and Hong Peng. Adaptive prototype and relation-aware distillation for multimodal sentiment analysis with incomplete data. *Neurocomputing*, 682:133397, 2026. doi: 10.1016/j.neucom.2026.133397.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656.

Nan Wang and Qi Wang. Modality-uncertainty-aware knowledge distillation framework for multimodal sentiment analysis. *Complex & Intelligent Systems*, 12:14, 2026. doi: 10.1007/s40747-025-02135-w.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21501–21509, 2025. doi: 10.1609/aaai.v39i20.35452.

Torsten Wörtwein, Nicholas B. Allen, Jeffrey F. Cohn, and Louis-Philippe Morency. SMURF: Statistical modality uniqueness and redundancy factorization. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pp. 339–349. Association for Computing Machinery, 2024. doi: 10.1145/3678957.3685716.

Zhonghao Xi, Bengong Yu, Shuping Zhao, and Shanlin Yang. Bridging the gaps: A hierarchical distillation approach to multimodal sentiment analysis with missing modalities. *Information Fusion*, 127:103838, 2026. doi: 10.1016/j.inffus.2025.103838.

Dongqing Xie, Yonghuang Wu, Zisheng Ai, Jun Min, Zhencun Jiang, Shaojin Geng, and Lei Wang. CCSD: Cross-modal compositional self-distillation for robust brain tumor segmentation with missing modalities. *Knowledge-Based Systems*, 346:116136, 2026. doi: 10.1016/j.knosys.2026.116136.

Chunlin Xu, Erbing Li, Huihui Li, Xiaoyong Liu, Weiqi Chen, and Jianhua Guo. Enhancing multimodal sentiment analysis via pairwise emotional correlation distillation and information bottleneck. *Neurocomputing*, 675:132653, 2026. doi: 10.1016/j.neucom.2026.132653.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jiwoong Yang, Haejun Chung, and Ikbeom Jang. Hierarchical mutual distillation for multi-view fusion: Learning from all possible view combinations. *Pattern Recognition*, 178:113432, 2026a. doi: 10.1016/j.patcog.2026.113432.

Zequn Yang, Hongfa Wang, and Di Hu. Efficient quantification of multimodal interaction at sample level. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71302–71317. PMLR, 2025b.

Zequn Yang, Yake Wei, Haotian Ni, Zhihao Xu, and Di Hu. Information-theoretic decomposition for multimodal interaction learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30278–30287, 2026b.

Mingyang Ye, Ming Yin, and Yi Guo. Multi-level relation-aware knowledge distillation with hierarchical fusion for incomplete multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 17(3):3521–3534, 2026. doi: 10.1109/TAFFC.2026.3687649.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 2016. doi: 10.1109/MIS.2016.94.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208.

Tiantai Zhai, Yan Zhuang, Fuji Ren, Jiawen Deng, and Liang Luo. ReNoRD: Learning from relations under noisy pseudo labels via relational distillation for multimodal sentiment. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 1355–1364. Association for Computing Machinery, 2026a. doi: 10.1145/3805622.3810693.

Yiqiao Zhai, Qiuxia Yang, Chengchao Wang, Lianmin Zhou, Jue Feng, Fanghong Hu, Zhengpeng Zhao, and Yuanyuan Pu. Hybrid prompt learning and multilevel knowledge distillation for multimodal sentiment analysis with missing modalities. *Expert Systems with Applications*, 315:131746, 2026b. doi: 10.1016/j.eswa.2026.131746.

Kaiwen Zheng, Junchen Fu, Wenhao Deng, Hu Han, Joemon M. Jose, and Xuri Ge. Do we really need multimodal emotion language models larger than 1B parameters?, 2026. Accepted at ACM Multimedia 2026.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. 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*, pp. 4626–4636, 2025. doi: 10.1109/ICCV51701.2025.00440.