From Coarse to Fine: A Multi-Granularity Multimodal Framework for Teacher Sentiment Analysis

Zhiyi Duan[1], Xiangren Wang[1], Jiangshan Guan[1], Bing Jia[1]*, Qianli Xing[2]*
[1]Department of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia, China
[2]College of Computer Science and Technology, Jilin University, Changchun, Jilin, China
duanzy@imu.edu.cn, wangxr@mail.imu.edu.cn, 32409140@mail.imu.edu.cn,
jiabing@imu.edu.cn, qianlixing@jlu.edu.cn

Abstract

Teacher sentiment analysis is pivotal for understanding instructional dynamics, yet it remains challenging because classroom expressions are professionally regulated performances rather than spontaneous outbursts. However, existing approaches typically treat sentiment as a static, monolithic label, failing to capture this structured heterogeneity. To effectively model this complexity, we decompose teacher sentiment into three granularities: coarse-level performativity, medium-level intra-class heterogeneity, and fine-level cross-modal complementarity. Guided by this perspective, we propose CF-TSA, a coarse-to-fine multimodal framework. Specifically, we employ CLS-guided cross-modal attention to recover effective signals from regulated displays (coarse-level), thresholded substyle discovery to identify latent pedagogical styles (medium-level), and substyle-aware contrastive learning to align dynamic multimodal cue compositions (fine-level). Experiments on T-MED and CMU-MOSEI demonstrate that CF-TSA consistently outperforms state-of-the-art baselines, validating the effectiveness of the coarse-to-fine perspective and the hierarchical modeling.

1 Introduction

Modeling teachers' sentiment in authentic classroom settings is theoretically important and practically impactful, as teachers' sentiments serve as a critical lever for instructional decisions, classroom climate, and student engagement (Frenzel et al., 2018). Despite its significance, reliable teacher sentiment analysis (TSA) remains an open challenge.

Real-world classrooms are complex environments where teachers' expressions are not spontaneous outbursts but structured phenomena deeply shaped by professional roles, pedagogical goals, and diverse signaling patterns (Wang and Frenzel, 2025). Despite this intricate reality, current methodologies fall short in two distinct ways. On one hand, domain-specific research often relies on single-modality signals (e.g., audio only) (Li, 2022; Cai et al., 2023; He et al., 2024), failing to capture the full spectrum of multimodal evidence required to decode such complex signals. On the other hand, while generic multimodal models fuse heterogeneous signals effectively, they typically treat sentiment as a static, monolithic label (Jin et al., 2024; Zhou et al., 2025; Zhang and Tan, 2025; Leng and Yan, 2025). Applying these models directly to education is insufficient because they lack the mechanisms to handle the specific gap between internal states and external performances inherent to teaching (Hou et al., 2024).

To effectively model such complexity, we argue that teacher sentiment should be decomposed across three distinct granularities, ranging from macroscopic professional constraints to microscopic signal compositions. At the coarse granularity, teacher sentiment is governed by performativity and professional regulation. Unlike spontaneous daily sentiments, a teacher's expression is a regulated "performance" guided by display rules. Teachers often mask their internal states, such as suppressing anger to maintain order or feigning enthusiasm to motivate students. At the medium granularity, sentiments manifest as diverse intra-class styles (Pekrun, 2006). Driven by varying instructional intents, a single sentiment label functions as a super-category containing multiple stable substyles. For instance, "Anger" in a classroom may appear as a high-energy disciplinary shout or a low-energy stern interrogation. At the fine granularity, expression relies on compositional complementarity with varying modality weights (Li et al., 2025). At the signal level, sentiment cues are dynamically composed rather than parallel. In one moment, the instructional intent may be carried primarily by

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 8861–8876

the semantic weight of text; in another, prosodic variation or visual gestures serve as the dominant evidence.

To operationalize this theoretical view, we propose CF-TSA, a framework that progressively refines sentiment representation from coarse to fine granularities. Specifically, to address professional performativity at the coarse level, we introduce a CLS-guided cross-modal attention module that treats the global token of one modality as a query to retrieve evidence from others. Subsequently, to model intra-class diversity at the medium level, we design a thresholded substyle discovery module with stability gating, which dynamically clusters samples within each sentiment class. Finally, to capture compositional complementarity at the fine level, we employ substyle-aware contrastive learning with prototype consolidation, optimizing the feature space to reflect the nuanced weight distribution of real-world classroom signals.

Our contributions are threefold:

- We systematically abstract the complexity of TSA into a hierarchical problem set: *coarse-level* signal recovery to address professional performativity, *medium-level* latent variable discovery to capture intra-class style heterogeneity, and *fine-level* compositional alignment to model contextual complementarity.
- We propose CF-TSA, a multi-granularity multimodal framework, which translates the above formulation into a concrete solution by CLS-guided cross-modal attention, thresholded substyle discovery, and substyle-aware contrastive objectives.
- Extensive experiments on the teacher-domain dataset T-MED and the general benchmark CMU-MOSEI demonstrate that CF-TSA achieves state-of-the-art performance, validating both the effectiveness of our model and the generalization capability of the proposed coarse-to-fine perspective.

2 Related Work

2.1 Teacher Sentiment Analysis

Teachers' sentiment expression in the classroom is constrained by instructional goals and professional display rules, and therefore exhibits the multi-granularity differences we mentioned. Existing studies provide empirical evidence for these properties: in classroom scenarios, prosodic cues are often more crucial (Zhao et al., 2024); classroom events may also display structured segmental characteristics, which supports explicitly incorporating interaction structure into the modeling process (Zheng et al., 2024).

At the modeling level, EDSN disentangles cross-modal shared information from modality-specific discriminative information, thereby preserving cross-modal consistency and unimodal cues in teacher sentiment analysis (Cai et al., 2025). More recent teacher-facing models further emphasize audio-centric modeling and adaptation to classroom settings (Duan et al., 2025). Although these studies align more closely with classroom constraints, they still remain at label-level learning and struggle to effectively model, at a deeper level, the distinctive properties inherent in teacher sentiment.

2.2 General Multimodal Affect Modeling

In contrast, modeling in general domains has gradually shifted toward interactive modeling or disentanglement (Zadeh et al., 2017, 2018a; Liu et al., 2018; Tsai et al., 2019; Wang et al., 2024), and has increasingly recognized that authentic affect is often difficult to describe accurately with a single label (Yang et al., 2023; Wang et al., 2025). However, existing approaches typically handle inconsistency, uncertainty, and missingness in isolation. They are also not explicitly coupled with mechanisms capable of discovering reliable substyles, which makes it difficult to systematically model within sentiment heterogeneity. So there remains a lack of paradigms that can be directly transferred to the teacher domain.

3 Problem Formulation: A Coarse-to-Fine Perspective

Let $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$ denote a multimodal dataset, where $X_i = \{X_i^T, X_i^A, X_i^V\}$ represents the text, audio, and visual signals, and $y_i \in \mathcal{Y}$ is the discrete sentiment label. Standard approaches typically aim to learn a direct mapping $P(y|X)$. However, guided by our theoretical framework, we model teacher sentiment analysis as a hierarchical process, decomposing the problem into three granularities:

Performativity as Signal Recovery (Coarse Granularity). We posit that the observed signal X is a regulated version of the internal affective state, distorted by a professional display function

$\mathcal{P}$. Consequently, the initial objective is not direct classification, but rather recovering the effective sentimental cues $\tilde{X}$ by retrieving complementary evidence across modalities. This can be formulated as a recovery function:

$$\tilde{X} = \text{Recover}(X_{\text{anchor}}, X_{\text{context}}) \tag{1}$$

where one modality serves as the structural anchor to retrieve evidence from the context of others.

Heterogeneity as Latent Variable Discovery (Medium Granularity). We assume the conditional distribution $P(X|y)$ is multi-modal. A single label y does not map to a centroid but generates data via a latent substyle variable $s \in \mathcal{S}_y$, where $\mathcal{S}_y$ is the set of stable substyles for class y. The objective is to identify this latent structure:

$$s^* = \arg\max_{s \in \mathcal{S}_y} P(s|\tilde{X}, y) \tag{2}$$

This necessitates a mechanism to discover and assign samples to these latent partitions.

Complementarity as Compositional Alignment (Fine Granularity). Finally, the prediction of y depends on the fine-grained composition of features weighted by the specific substyle s. We seek a feature space where samples are aligned by their specific compositional pattern:

$$\mathcal{L}_{\text{fine}} = \mathbb{E}[\text{sim}(\tilde{X}, \mathbf{p}_{y,s})] \tag{3}$$

where $\mathbf{p}_{y,s}$ denotes the prototype for substyle s of class y, and $\text{sim}(\cdot)$ is a similarity metric.

In our proposed framework **CF-TSA**, we mechanistically instantiate these theoretical targets. Specifically, the recovery process $\tilde{X}$ corresponds to the fused representation z obtained via *CLS-guided cross-modal attention*; the latent variable s is solved via *thresholded substyle discovery*; and the alignment target is optimized via *substyle-aware contrastive learning*.

4 Method

In this section, we propose **CF-TSA**, a multi-granularity framework illustrated in Figure 1. To address the three challenges mentioned above, **CF-TSA** progressively utilizes CLS-guided cross-modal attention to recover effective signals from coarse-level performativity (Sec.4.2); applies thresholded substyle discovery to capture medium-level intra-class heterogeneity (Sec.4.3); and leverages substyle-aware contrastive learning to align fine-level multimodal compositions (Sec.4.4).

4.1 Multimodal Feature Extraction Module

For sample i, the inputs are transcript x_i^T, audio x_i^A, and video x_i^V. We extract modality-specific features with RoBERTa-base for text (Liu et al., 2019), FACET for vision (Stöckli et al., 2018), and COVAREP (general-domain) or HuBERT-base (teacher-domain) for audio (Degottex et al., 2014; Hsu et al., 2021). Based on these features, our module produces for each modality (i) a compact global representation and (ii) a token/frame-level structure representation, together with a validity mask to handle variable-length sequences and padding.

Concretely, the text outputs are the global vector g_i^T and the structure sequence $H_{i,\text{str}}^T$ with token mask m_i^T. For audio and vision, the outputs are global vectors g_i^A and g_i^V, structure sequences H_i^A and H_i^V, and corresponding frame masks m_i^A and m_i^V. All formulation details, e.g., [CLS] token handling, mask definitions, and the aggregation procedure, are provided in Appendix A.

4.2 Coarse Level: Performativity-aware Signal Recovery

To implement the signal recovery function $\tilde{X} = \text{Recover}(X_{\text{anchor}}, X_{\text{context}})$ defined in Eq. (1), we propose a **CLS-guided Cross-modal Attention Module**. This design addresses professional performativity: when a teacher's expression in one modality is regulated (e.g., maintaining a calm face while feeling angry), the model uses a structural anchor to actively retrieve complementary evidence from other modalities, thereby "seeing through" the regulation. We obtain an anchor vector a_i^m and an evidence bank S_i^m for each modality $m \in \{T, A, V\}$ via modality-specific projections; details are provided in Appendix B.

4.2.1 CLS Guided Cross Attention.

We present the text-query branch as an example: the text anchor a_i^T serves as the query, while the audio and vision evidence banks are concatenated as the context for key and value:

$$\tilde{S}_i^T = \left[S_i^A; S_i^V\right], \qquad \tilde{m}_i^T = \left[m_i^A; m_i^V\right], \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation along the sequence dimension and $\tilde{m}_i^T$ matches the length of $\tilde{S}_i^T$. We update the anchor by cross attention over the concatenated context, followed by a feed-forward block:

$$\begin{aligned} \text{XAtt}(q, S;\, m) &= \text{LN}\Big(q + \text{Att}(q, S;\, m)\Big), \\ \text{Att}(q, S;\, m) &= \text{MHA}(q, S, S;\, m), \end{aligned} \tag{5}$$

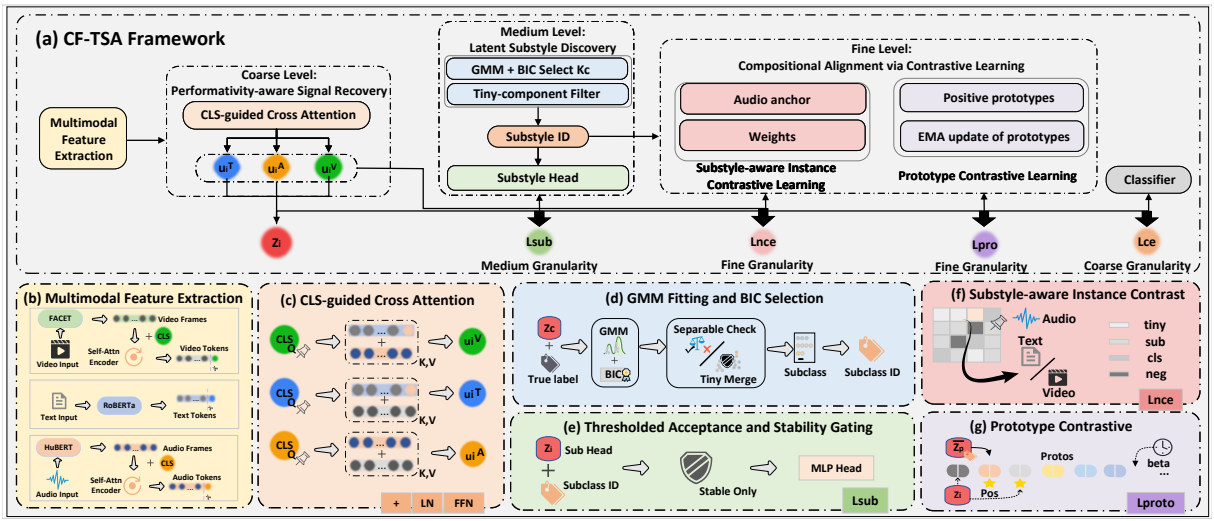


Figure 1: Overview of the CF-TSA Framework. The architecture processes multimodal inputs through a hierarchical pipeline designed to address three levels of complexity: (1) Coarse Level: The CLS-guided Cross-Modal Attention module recovers effective signals from regulated expressions to handle professional performativity; (2) Medium Level: The Thresholded Substyle Discovery module identifies latent pedagogical substyles to model intra-class heterogeneity; and (3) Fine Level: The Substyle-aware Contrastive Learning module aligns dynamic cross-modal cue compositions to capture contextual complementarity.

$$\mathrm{Post}(q) = \mathrm{LN}\Big(q + \mathrm{FFN}(q)\Big), \tag{6}$$

where $\mathrm{MHA}(\cdot)$ is multi-head attention with an attention mask, $\mathrm{FFN}(\cdot)$ is a feed forward network, and $\mathrm{LN}(\cdot)$ is layer normalization. The cross-informed text anchor is:

$$u_i^T = \mathrm{Post}\Big(\mathrm{XAtt}(a_i^T, \tilde{S}_i^T; \tilde{m}_i^T)\Big). \tag{7}$$

We employ the same cross-attention process for the audio-query and vision-query branches to obtain u_i^A and u_i^V, and the detailed instantiations are provided in Appendix C. We concatenate the updated anchors as the fused representation:

$$z_i = \left[u_i^T; u_i^A; u_i^V\right] \in \mathbb{R}^{3d}. \tag{8}$$

The final fused representation is the concatenation of these recovered signals: $z_i = [u_i^T; u_i^A; u_i^V]$, which serves as the instantiation of the recovered cue $\tilde{X}$ in our formulation.

4.3 Medium Level: Latent Substyle Discovery

Having recovered the effective signals z_i, we next aim to identify the latent substyle structure s^* as formulated in Eq. (2). This addresses intra-class style heterogeneity, where a single sentiment label y acts as a super-category governing multiple pedagogical realizations.

To explicitly model this multi-modal conditional distribution $P(X|y)$, we employ a **Thresholded Substyle Discovery Module**. This module addresses intra-class style heterogeneity by discovering recurrent substyles within each sentiment class. To obtain stable substyle assignments, we use a thresholded procedure to avoid over-splitting and apply a size-based filter to gate unreliable tiny components. For each class c, we collect fused representations:

$$\mathcal{Z}_c = \{z_i \mid y_i = c\}, \tag{9}$$

where $y_i \in \{1, \ldots, C\}$ is the class label and C is the number of classes.

4.3.1 GMM Fitting and BIC Selection

We fit Gaussian mixture models on $\mathcal{Z}_c$ and select the number of components K_c via BIC (Schwarz, 1978). We define the relative gain:

$$\Delta_c = \mathrm{BIC}_c(1) - \mathrm{BIC}_c(K_c). \tag{10}$$

Let $\{\mu_{c,k}\}_{k=1}^{K_c}$ denote the component means under the selected model. We define an angular separability statistic:

$$\delta_c = \min_{k \neq k'} \angle(\mu_{c,k}, \mu_{c,k'}), \tag{11}$$

where $\angle(u, v) = \arccos\left(\frac{u^\top v}{\|u\|\,\|v\|}\right)$.

4.3.2 Thresholded Acceptance and Stability Gating

We apply two thresholds: a BIC gain constraint $\Delta_c \geq \gamma$ and an angular constraint $\delta_c \geq \delta_{\text{ang}}$. If

either constraint fails, we revert to a single component assignment ($K_c \leftarrow 1$) for class c.

Let $\pi(i)$ denote the component assignment for sample i under the (possibly reverted) model of class y_i. We define component subsets:

$$\mathcal{Z}_{c,k} = \{z_i \mid y_i = c,\ \pi(i) = k\}, \tag{12}$$

where $k \in \{1, \ldots, K_c\}$. We treat components with size smaller than $n_{\min}$ as tiny and define:

$$\mathcal{H}_c = \left\{ i \;\middle|\; y_i = c,\ \left|\mathcal{Z}_{c,\pi(i)}\right| < n_{\min} \right\}. \tag{13}$$

We define a stability gate:

$$r_i = \mathbb{I}\big[i \notin \mathcal{H}_{y_i}\big], \tag{14}$$

where $r_i = 1$ indicates a stable sample and $r_i = 0$ indicates a sample assigned to a tiny component. For stable samples, we define the substyle index:

$$s_i = \pi(i) \quad \text{when } r_i = 1. \tag{15}$$

4.3.3 Substyle Auxiliary Supervision Loss

With the stable substyle assignments obtained above, we introduce an auxiliary substyle head g_ϕ to predict s_i within the stable substyle set of class y_i. We denote the stable substyle index set for class c as:

$$\mathcal{S}_c = \left\{ k \;\middle|\; \left|\mathcal{Z}_{c,k}\right| \geq n_{\min} \right\}. \tag{16}$$

Then the conditional prediction is:

$$P_\phi(s \mid z_i, y_i) = \mathrm{Softmax}\big(g_\phi(z_i, y_i)\big)_s, \quad s \in \mathcal{S}_{y_i}. \tag{17}$$

Let $\mathcal{B}_{\text{sub}}$ denote the set of samples used for substyle supervision (i.e., stable samples with $r_i = 1$). The loss is:

$$\mathcal{L}_{\text{sub}} = -\frac{1}{|\mathcal{B}_{\text{sub}}|} \sum_{i \in \mathcal{B}_{\text{sub}}} \log P_\phi(s_i \mid z_i, y_i). \tag{18}$$

4.4 Fine Level: Compositional Alignment via Contrastive Learning

Since different substyles rely on distinct combinations of modality weights, simple class-level alignment is insufficient. To optimize the fine-grained alignment target $\mathcal{L}_{\text{fine}}$ in Eq. (3), we introduce a **Substyle-aware Contrastive Learning Module**.

4.4.1 Substyle-aware Instance Contrastive Learning

We define cosine similarity and a temperature:

$$s(u, v) = \frac{u^\top v}{\|u\|\,\|v\|}, \qquad \tau > 0. \tag{19}$$

Substyle-aware weights are specified as:

$$\omega_{ij} = \begin{cases} \omega_{\text{tiny}}, & r_j = 0, \\ \omega_{\text{sub}}, & y_i = y_j,\ r_i = r_j = 1,\ s_i = s_j, \\ \omega_{\text{cls}}, & y_i = y_j,\ r_i = r_j = 1,\ s_i \neq s_j, \\ \omega_{\text{neg}}, & y_i \neq y_j, \end{cases} \tag{20}$$

$$\omega_{\text{tiny}} < \omega_{\text{sub}} < \omega_{\text{cls}} < \omega_{\text{neg}}. \tag{21}$$

Using audio as the anchor, we contrast the audio-updated representation against text and vision:

$$\ell_i^{A\to T} = -\log \frac{\exp\big(s(u_i^A, u_i^T)/\tau\big)}{\sum_{j=1}^{B} \omega_{ij} \exp\big(s(u_i^A, u_j^T)/\tau\big)}, \tag{22}$$

$$\ell_i^{A\to V} = -\log \frac{\exp\big(s(u_i^A, u_i^V)/\tau\big)}{\sum_{j=1}^{B} \omega_{ij} \exp\big(s(u_i^A, u_j^V)/\tau\big)}, \tag{23}$$

where B is the mini-batch size. We gate the loss to stable samples:

$$\mathcal{L}_{\text{nce}} = \frac{1}{\sum_{i=1}^{B} r_i} \sum_{i=1}^{B} r_i \Big(\ell_i^{A\to T} + \ell_i^{A\to V}\Big). \tag{24}$$

We employ the same instance-contrast construction when selecting a different anchor modality, and the detailed text-anchored form is provided in Appendix D (Vaswani et al., 2017).

4.4.2 Prototype Contrastive Learning

We keep prototypes for classes and stable substyles:

$$\mathcal{P} = \{p_c\}_{c=1}^{C} \cup \{p_{c,k} \mid k \in \mathcal{S}_c\}, \tag{25}$$

where p_c is the prototype for class c, $p_{c,k}$ is the prototype for stable substyle k of class c, and $\mathcal{S}_c$ is the stable substyle index set for class c.

For a stable sample, the positive prototype set is:

$$\mathcal{P}^+(i) = \{p_{y_i},\ p_{y_i,s_i}\} \quad \text{when } r_i = 1. \tag{26}$$

Each prototype is assigned a nonnegative weight $\eta(p) \geq 0$, and we define:

$$\psi_{i,p} = \exp(\eta(p)\, s(z_i, p)/\tau)\,. \tag{27}$$

The prototype contrast loss is:

$$\ell_i^{\text{proto}} = -\log \frac{\sum_{p\in\mathcal{P}^+(i)} \psi_{i,p}}{\sum_{p\in\mathcal{P}} \psi_{i,p}}, \tag{28}$$

$$\mathcal{L}_{\text{proto}} = \frac{1}{\sum_{i=1}^{B} r_i} \sum_{i=1}^{B} r_i\, \ell_i^{\text{proto}}. \tag{29}$$

For EMA updates, we define index sets for each prototype (Caron et al., 2020):

$$\begin{aligned} \mathcal{I}(p_c) &= \{i \mid r_i = 1,\ y_i = c\}, \\ \mathcal{I}(p_{c,k}) &= \{i \mid r_i = 1,\ y_i = c,\ s_i = k\}. \end{aligned} \tag{30}$$

The batch mean is computed by:

$$\bar{z}_p = \text{Mean}\big(\{z_i \mid i \in \mathcal{I}(p)\}\big), \tag{31}$$

and update prototypes with EMA:

$$p \leftarrow \beta p + (1-\beta)\bar{z}_p, \qquad \beta \in [0,1). \tag{32}$$

4.5 Classification Module

We classify the fused representation z_i with an angular margin softmax classifier:

$$\cos(\theta_{i,c}) = \frac{w_c^\top z_i}{\|w_c\|\,\|z_i\|}, \tag{33}$$

where w_c is the classifier weight vector for class c and $\theta_{i,c}$ is the angle between z_i and w_c. We define the logit with scaling α and angular margin m_{arc}:

$$\ell_{i,c} = \alpha \cdot \cos\big(\theta_{i,c} + m_{\text{arc}} \cdot \mathbb{I}[c = y_i]\big). \tag{34}$$

The per-sample cross entropy is:

$$\ell_i^{\text{ce}} = -\log \frac{\exp(\ell_{i,y_i})}{\sum_{c'=1}^{C} \exp(\ell_{i,c'})}. \tag{35}$$

4.6 Optimization Objective

We jointly optimize the classification loss, substyle supervision, instance contrast, and prototype contrast. Specifically, we define the classification loss as the mini-batch averaged cross entropy:

$$\mathcal{L}_{\text{ce}} = \frac{1}{B} \sum_{i=1}^{B} \ell_i^{\text{ce}}, \tag{36}$$

and optimize the overall objective:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda_{\text{sub}}\mathcal{L}_{\text{sub}} + \lambda_{\text{nce}}\mathcal{L}_{\text{nce}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}}, \tag{37}$$

where λ_{sub}, λ_{nce}, and λ_{proto} are weighting hyperparameters.

5 Experiments

Datasets. We evaluate on the teacher-domain sentiment dataset T-MED (Duan et al., 2025) (14,938 multimodal classroom instances) and the widely used general-domain benchmark CMU-MOSEI (Zadeh et al., 2018b) (22,856 instances), following the official splits, with dataset details deferred to Appendix E; all methods are trained under a unified protocol with consistent preprocessing; additional details on metrics and implementation are provided in Appendix F. Parameter sensitivity experiments are presented in Appendix G.

Baselines. General domain baselines include TFN, MFN, LMF, MulT, MISA(Hazarika et al., 2020), ConFEDE, DFMU(Tang et al., 2025), DLF, and MFMB-Net(Tao et al., 2025). We additionally include teacher oriented baselines EDSN, ST-SER, and AAM-TSA. Implementation details and brief descriptions are provided in Appendix H.

5.1 Main Results

Teacher Domain (T-MED). CF-TSA achieves state-of-the-art performance, peaking at 86.11 Acc and 85.31 W-F1 under the $T{+}A$ setting. This surpasses the strongest baseline AAM-TSA by +0.94 Acc, verifying the efficacy of our coarse-to-fine design. Interestingly, introducing the visual modality ($T{+}A{+}V$) yields slightly lower performance than $T{+}A$. This aligns with classroom realities where visual streams are often dominated by non-affective motion (e.g., board writing) and environmental clutter, whereas prosody offers more concentrated sentimental evidence. CF-TSA effectively exploits this by prioritizing high-quality audio-text signals.

General Domain (CMU-MOSEI). Although motivated by educational constraints, CF-TSA demonstrates strong generalization. It outperforms the strongest baseline DFMU by +1.09 Acc in the binary setting (with 0). Even in the high-ceiling binary setting (w/o neutral), CF-TSA maintains the lead. Notably, on the challenging 7-class task, it achieves 55.16 Acc7, exceeding the previous best by +0.30. These results confirm that our framework's ability to recover regulated signals and align compositional cues is beneficial across diverse multimodal contexts.

5.2 Deconstructing the Coarse-to-Fine Framework

To provide a comprehensive analysis of the proposed framework, we systematically validate CF-

Model	Modalities			T-MED		CMU-MOSEI				
	T	A	V	Acc (%)	W-F1 (%)	Acc (Bin)	W-F1 (Bin)	Acc (w/o 0)	W-F1 (w/o 0)	Acc7
TFN (2017)	✓	✓	✓	75.23	74.15	78.50	78.96	81.89	81.74	51.60
MFN (2018)	✓	✓	✓	77.84	76.71	78.94	79.55	82.86	82.85	51.34
LMF (2018)	✓	✓	✓	76.72	75.83	80.54	80.94	83.48	83.36	51.59
MulT (2019)	✓	✓	✓	79.20	78.54	81.15	81.56	84.63	84.52	52.84
MISA (2020)	✓	✓	✓	78.93	78.26	83.60	83.80	85.50	85.30	52.20
ConFEDE (2023)	✓	✓	✓	80.19	79.36	81.65	82.17	85.82	85.83	54.86
ST-SER (2024)	✓	✓	×	76.32	75.91	78.21	78.56	81.35	81.17	50.02
MFMB-Net (2025)	✓	✓	✓	80.91	79.81	84.70	85.00	85.10	85.10	54.20
DFMU (2025)	✓	✓	✓	81.70	80.85	85.65	85.61	87.59	87.62	54.50
DLF (2025)	✓	✓	✓	81.30	80.45	83.69	83.90	86.02	86.10	54.17
EDSN (2025)	✓	✓	✓	82.40	81.70	83.71	83.92	86.12	86.03	53.73
AAM-TSA (2025)	✓	✓	×	85.17	84.34	82.51	82.27	83.74	83.56	52.97
AAM-TSA (2025)	✓	✓	✓	84.57	83.78	83.26	83.57	84.32	84.21	53.59
CF-TSA	✓	✓	×	**86.11**	**85.31**	86.09	86.09	87.42	87.45	54.54
CF-TSA	✓	✓	✓	85.44	84.21	**86.74**	**86.61**	**87.73**	**87.73**	**55.16**

Table 1: Results on the T-MED and CMU-MOSEI datasets. T/A/V denote text/audio/vision. Metrics of T-MED: Acc and W-F1. Metrics of CMU-MOSEI: Acc/W-F1 (Bin), Acc/W-F1 (w/o 0), and Acc7. Best results are in bold.

TSA through grouped ablation studies. By deconstructing the framework according to the three granularities, we isolate the specific contribution of each module from coarse-level signal recovery to fine-level compositional alignment by incrementally adding components to the minimal baseline VAR-NULL. We report results on T-MED ($T+A$) in Table 2, quantifying the performance gains to verify the necessity and importance of each component, while also analyzing the statistical properties of the learned structures to interpret the underlying pedagogical heterogeneity.

Coarse Level: Signal Recovery via Anchoring. By comparing anchoring strategies, we observe that the audio-anchored variant (VAR-A-only) achieves a decisive gain of +8.33 Acc and +12.31 W-F1 over the text-anchored baseline (VAR-T-only). This empirical evidence supports our coarse-level assumption: in regulated classroom settings, prosodic cues are less susceptible to masking than semantic content, making them the optimal anchor for recovering effective sentimental signals.

Medium Level: Heterogeneity via Thresholded Discovery. We examine the mechanism for modeling intra-class heterogeneity by analyzing the discovery strategies. As shown in Groups 2 and 3, replacing standard K-Means (KM) with our Thresholded Discovery (DG) yields consistent gains regardless of the alignment objective. Specifically, VAR-DG-IC outperforms VAR-KM-IC by +0.59 Acc, and similarly, VAR-DG-SC surpasses VAR-KM-SC. This robustness indicates that simply partitioning data is insufficient.

The necessity of our design is further supported by the substyle statistics in Table 3 (more details can be found in Appendix I). The method discovers an average of 4.25 stable substyles per category, confirming the one-label-many-styles hypothesis. Crucially, the high variance in substyle sizes, ranging from 42 to 712, highlights the prevalence of unstable noise. By explicitly filtering these tiny components, our DG strategy successfully captures the reliable pedagogical structure, directly translating to the performance gains.

Fine Level: Complementarity via Compositional Alignment. Finally, we evaluate the alignment of multimodal compositions. We find that neither substyle supervision (SC) nor instance contrast (IC) is sufficient alone. However, their combination (VAR-DG-SCIC) yields a substantial boost, with +0.97 Acc over SC only, validating that fine-grained alignment requires jointly reinforcing within-substyle structure and instance discrimination. Furthermore, adding Prototype Consolidation (PR) provides a final gain of +0.35 Acc, proving that maintaining global prototypes helps stabilize recurring cross-modal cue compositions. More fine grained analyses are provided in Appendix J.

Variant	KM	DG	SC	IC	PR	Acc	W-F1
(G1) Coarse Level: Signal Recovery via Anchor Selection							
VAR-T-only	×	×	×	×	×	70.21	64.92
VAR-A-only	×	×	×	×	×	78.54	77.23
(G2) Medium Level: Discovery Mechanism (w/ Contrast)							
VAR-NULL	×	×	×	×	×	83.57	82.44
VAR-KM-IC	✓	×	×	✓	×	84.45	83.21
VAR-DG-IC	×	✓	×	✓	×	85.04	84.31
(G3) Medium Level: Discovery Mechanism (w/ Supervision)							
VAR-NULL	×	×	×	×	×	83.57	82.44
VAR-KM-SC	✓	×	✓	×	×	84.23	83.12
VAR-DG-SC	×	✓	✓	×	×	84.79	83.72
(G4) Fine Level: Compositional Alignment (SC vs. IC)							
VAR-NULL	×	×	×	×	×	83.57	82.44
VAR-DG-SC	×	✓	✓	×	×	84.79	83.72
VAR-DG-IC	×	✓	×	✓	×	85.04	84.31
VAR-DG-SCIC	×	✓	✓	✓	×	85.76	84.93
(G5) Fine Level: Prototype Consolidation							
VAR-DG-SCIC	×	✓	✓	✓	×	85.76	84.93
VAR-FULL	×	✓	✓	✓	✓	**86.11**	**85.31**

Table 2: Results of Deconstructing the Coarse-to-Fine framework on T-MED ($T+A$). KM: KMeans based substyle discovery; DG: thresholded discovery; SC: substyle-aware classification supervision; IC: substyle-aware instance contrast learning; PR: prototype consolidation for cross-modal patterns.

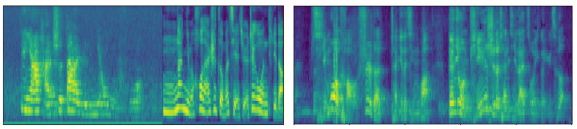


Figure 2: Visualization of "Expectation": One Label, Distinct Substyles.

5.3 Case Study: Decoding Sentiment Granularities

To complement our quantitative results, we present a qualitative case study to illustrate how CF-TSA decodes the complexity of teacher sentiment across three granularities. Figure 2 visualizes two instances labeled as Expectation, which share a high-level goal of advancing instruction but diverge significantly in their realization.

Transcript a_1: "The voltage is still above 0.5V, so how did you solve this problem afterward?"

Transcript a_2: "Regarding final exam results, let's make a prediction based on usual grades."

Although both instances are annotated with the same label (y = Expectation), our framework decomposes their differences as follows.

Statistic	**Value**
Number of categories (C)	8
Substyles per category (min/mean/max)	2 / 4.25 / 6
Substyles per category (p10 / p50 / p90)	2.0 / 4.0 / 6.0
Total number of substyles	34
Substyle size (p10 / p50 / p90)	42 / 196 / 712
Number of prototypes (class / substyle)	8 / 34

Table 3: Statistics of discovered stable substyles and maintained prototypes on T-MED under $T+A$.

Coarse Level (Signal Recovery): Teacher expressions are professionally regulated. As shown in Figure 2, while both texts convey instructional intent, audio serves as the structural anchor: a_1 uses segmented prosody for interaction, whereas a_2 uses continuous delivery for framing. Our CLS-Guided Cross-Modal Attention leverages these cues to recover effective signals.

Medium Level (Substyle Discovery): The analysis confirms that a single sentiment label functions as a super-category containing multiple latent substyles. a_1 represents a "Response-seeking" style, while a_2 reflects a "Task-framing" style. Our Thresholded Substyle Discovery explicitly models this heterogeneity rather than merging them into a single centroid.

Fine Level (Compositional Alignment): Finally, the visualization reveals that sentiment expression relies on specific cross-modal compositions. a_1 pairs interrogative text with turn-taking gaps, whereas a_2 aligns declarative text with explanatory flow. Our substyle-aware contrastive learning effectively captures these context-dependent patterns. Further analysis with more cases and expanded details is presented in Appendix K.

6 Conclusion

In this work, we argue that teacher sentiment analysis requires moving beyond static, monolithic labels to model the structured complexity of classroom expressions. To this end, we systematically abstract the challenge into three sub-tasks: signal recovery for performativity, latent variable discovery for heterogeneity, and compositional alignment for complementarity. Correspondingly, we propose CF-TSA, a framework that translates these theoretical targets into concrete solutions via CLS-guided cross-modal attention, thresholded substyle discovery, and substyle-aware contrastive learning.

Comprehensive evaluations, including spanning comparative benchmarks, rigorous ablations, and fine-grained case studies, demonstrate that CF-TSA not only achieves state-of-the-art performance on T-MED and CMU-MOSEI but also exhibits superior robustness against domain-specific noise. Ultimately, our work establishes a robust framework for decoding professional affect, advancing the interdisciplinary convergence of educational theory and multimodal AI.

Limitations

Our current model and experiments focus on three modalities (text, audio, and video), and do not yet incorporate richer pedagogical context signals that could support more fine-grained and interpretable substyle modeling. Classroom video also presents substantial domain-specific clutter and non-affective motion, so stronger subject-centric and noise-robust spatiotemporal representations remain necessary for reliable visual modeling. In addition, we have not systematically studied modality missingness in real-world deployment, and robustness under incomplete multimodal inputs is an important direction. Finally, real-world deployment raises privacy considerations and calls for broader validation across diverse cultural and regional contexts. In future work, we will further investigate these issues in greater depth.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Nos. 62567005 and 62406127), the Natural Science Foundation of Inner Mongolia Autonomous Region of China (No. 2025MS06004), and in part by the Program for Young Talents of Science and Technology in Universities of Inner Mongolia A. R. of China under Grant NJYT25011.

References

Ting Cai, Shengsong Wang, Jing Wang, Yu Xiong, and Long Liu. 2025. Emotion dual-space network based on common and discriminative features for multimodal teacher emotion recognition. *Frontiers of Digital Education*, 2(3):25.

Ting Cai, Shengsong Wang, Yu Xiong, and Xin Zhong. 2023. Exploiting adaptive adversarial transfer network for cross domain teacher's speech emotion recognition. In *International Conference on Computer Science and Education*, pages 202–213. Springer.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

Zhiyi Duan, Xiangren Wang, Hongyu Yuan, and Qianli Xing. 2025. Advancing multimodal teacher sentiment analysis: The large-scale t-med dataset & the effective aam-tsa model. *arXiv preprint arXiv:2512.20548*.

Anne C Frenzel, Betty Becker-Kurz, Reinhard Pekrun, Thomas Goetz, and Oliver Lüdtke. 2018. Emotion transmission in the classroom revisited: a reciprocal effects model of teacher and student enjoyment. *Journal of Educational Psychology*, 110(5):628.

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.

Yimin He, Xiaoyong Lu, Dan Sun, Tao Pan, Yuqing Qiu, and Jiahong Liu. 2024. Research on teacher classroom teaching speech emotion recognition based on lstm. In *2024 International Conference on Asian Language Processing (IALP)*, pages 326–331. IEEE.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and chatgpt. In *International conference on artificial intelligence in education*, pages 60–74. Springer.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Tao Jin, Wang Lin, Ye Wang, Linjun Li, Xize Cheng, and Zhou Zhao. 2024. Rethinking the multimodal correlation of multimodal sequential learning via generalizable attentional results alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5247–5265.

Jidong Leng and Qiang Yan. 2025. Eijl: Popularity prediction of social media advertisements based on multimodal emotional interaction and joint learning. *Data Intelligence*, 7(4):1129–1146.

Liqin Li. 2022. Emotion analysis method of teaching evaluation texts based on deep learning in big data environment. *Computational intelligence and neuroscience*, 2022(1):9909209.

Pei-Hsin Li, Diane Mayer, and Lars-Erik Malmberg. 2025. Are teachers' and students' emotions reciprocally transmitted in the classroom? *British Journal of Educational Psychology*, 95:S172–S193.

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, AmirAli Bagher 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 (Volume 1: Long Papers)*, pages 2247–2256.

Reinhard Pekrun. 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, 18(4):315–341.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Sabrina Stöckli, Michael Schulte-Mecklenbeck, Stefan Borer, and Andrea C Samson. 2018. Facial expression analysis with affdex and facet: A validation study. *Behavior research methods*, 50(4):1446–1460.

Chen Tang, Tingrui Shen, Xinrong Gong, Chong Zhao, and Tong Zhang. 2025. Dfmu: Distribution-based framework for modeling aleatoric uncertainty in multimodal sentiment analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 8250–8258.

Chuanqi Tao, Jiaming Li, Tianzi Zang, and Peng Gao. 2025. A multi-focus-driven multi-branch network for robust multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1547–1555.

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 conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Hui Wang and Anne C Frenzel. 2025. "exhaustive but effective": A multi-site study investigating the profiles of teachers' emotions and emotional labor. *Journal of School Psychology*, 110:101456.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188.

Yong Wang, Ningchuang Yang, Duoqian Miao, and Qiuyi Chen. 2024. Aspect-guided multi-graph convolutional networks for aspect-based sentiment analysis. *Data Intelligence*, 6(3):771–791.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Tao Zhang and Zhenhua Tan. 2025. Ecerc: evidence-cause attention network for multi-modal emotion recognition in conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2064–2077.

Gang Zhao, Yinan Zhang, and Jie Chu. 2024. A multimodal teacher speech emotion recognition method in the smart classroom. *Internet of Things*, 25:101069.

Qiuyu Zheng, Zengzhao Chen, Mengke Wang, Yawen Shi, Shaohui Chen, and Zhi Liu. 2024. Automated multimode teaching behavior analysis: A pipeline-based event segmentation and description. *IEEE Transactions on Learning Technologies*, 17:1677–1693.

Miao Zhou, Lina Yang, Thomas Wu, Dongnan Yang, and Xinru Zhang. 2025. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11366–11376.

A Details of Multimodal Feature Extraction

This appendix provides the complete formulations of the multimodal feature extraction module described in Section 4.1, including the definitions of global vectors, structure sequences, and validity masks for text, audio, and vision.

A.1 Text feature details

A pretrained language encoder produces token representations

$$H_i^T = \big[h_{i,0}^T, h_{i,1}^T, \ldots, h_{i,L_T}^T\big], \tag{38}$$

where L_T is the number of tokens and $h_{i,0}^T$ corresponds to the [CLS] token. We use the [CLS] vector as the text global representation:

$$g_i^T = h_{i,0}^T. \tag{39}$$

The text structure sequence excludes [CLS] and aligns to a token validity mask $m_i^T \in \{0,1\}^{L_T}$:

$$H_{i,\mathrm{str}}^T = \big[h_{i,1}^T, \ldots, h_{i,L_T}^T\big]. \tag{40}$$

A.2 Audio aggregation details

An acoustic feature extractor yields a frame sequence

$$U_i^A = \big[u_{i,1}^A, u_{i,2}^A, \ldots, u_{i,L_A}^A\big], \tag{41}$$

where L_A is the number of audio frames. We prepend a learnable audio [CLS] token e^A:

$$X_i^A = \big[e^A, U_i^A\big]. \tag{42}$$

Let $m_i^A \in \{0,1\}^{L_A}$ be the frame validity mask and define the extended mask

$$\bar{m}_i^A = \big[1, m_{i,1}^A, \ldots, m_{i,L_A}^A\big], \tag{43}$$

where the prepended [CLS] position is always valid. We aggregate temporal evidence with a lightweight Transformer encoder:

$$\begin{aligned} \bar{H}_i^A &= \mathrm{AggEnc}\big(X_i^A; \bar{m}_i^A\big) \\ &= \big[h_{i,0}^A, h_{i,1}^A, \ldots, h_{i,L_A}^A\big]. \end{aligned} \tag{44}$$

We define the audio global vector and structure sequence:

$$g_i^A = h_{i,0}^A, \qquad H_i^A = \big[h_{i,1}^A, \ldots, h_{i,L_A}^A\big]. \tag{45}$$

A.3 Vision aggregation details

A visual feature extractor yields a frame sequence

$$U_i^V = \big[u_{i,1}^V, u_{i,2}^V, \ldots, u_{i,L_V}^V\big], \tag{46}$$

where L_V is the number of visual frames.

We prepend a learnable vision [CLS] token e^V and define the extended mask:

$$\begin{aligned} X_i^V &= \big[e^V, U_i^V\big], \\ \bar{m}_i^V &= \big[1, m_{i,1}^V, \ldots, m_{i,L_V}^V\big], \end{aligned} \tag{47}$$

where $m_i^V \in \{0,1\}^{L_V}$ is the frame validity mask.

We aggregate temporal evidence:

$$\begin{aligned} \bar{H}_i^V &= \mathrm{AggEnc}\big(X_i^V; \bar{m}_i^V\big) \\ &= \big[h_{i,0}^V, h_{i,1}^V, \ldots, h_{i,L_V}^V\big]. \end{aligned} \tag{48}$$

We define the vision global vector and structure sequence:

$$g_i^V = h_{i,0}^V, \qquad H_i^V = \big[h_{i,1}^V, \ldots, h_{i,L_V}^V\big]. \tag{49}$$

B Modality-specific Projection to a Shared Space

This appendix provides the projection details for constructing the anchor vector a_i^m and the evidence bank S_i^m used in the CLS-guided cross-modal attention module (Section 4.2).

We project each modality into a shared d-dimensional space, separating a CLS anchor and a structure evidence bank:

$$a_i^m = W_g^m g_i^m, \quad m \in \{T, A, V\}, \tag{50}$$

$$\begin{aligned} S_i^T &= W_s^T H_{i,\mathrm{str}}^T, \\ S_i^A &= W_s^A H_i^A, \\ S_i^V &= W_s^V H_i^V, \end{aligned} \tag{51}$$

where W_g^m and W_s^m are modality-specific projection matrices, $a_i^m \in \mathbb{R}^d$ is the anchor vector, and S_i^m is the evidence bank.

C Other Cross-Modal Attention Instantiations

In Section 4.2, we detail the text-query branch as an example. Here we provide the audio-query and vision-query instantiations for completeness.

C.1 Audio as query

For the audio-query branch, the audio anchor a_i^A serves as the query, while text and vision evidence banks are concatenated as the context:

$$\tilde{S}_i^A = \big[S_i^T; S_i^V\big], \qquad \tilde{m}_i^A = \big[m_i^T; m_i^V\big]. \quad (52)$$

The cross-informed audio anchor is:

$$u_i^A = \mathrm{Post}\Big(\mathrm{XAtt}(a_i^A, \tilde{S}_i^A; \tilde{m}_i^A)\Big). \quad (53)$$

C.2 Vision as query

For the vision-query branch, the vision anchor a_i^V serves as the query, while text and audio evidence banks are concatenated as the context:

$$\tilde{S}_i^V = \big[S_i^T; S_i^A\big], \qquad \tilde{m}_i^V = \big[m_i^T; m_i^A\big]. \quad (54)$$

The cross-informed vision anchor is:

$$u_i^V = \mathrm{Post}\Big(\mathrm{XAtt}(a_i^V, \tilde{S}_i^V; \tilde{m}_i^V)\Big). \quad (55)$$

D Alternative Instance-Contrast Variant

In Section 4.4, we describe audio-anchored instance contrast for the $T{+}A{+}V$ configuration. We employ the same construction when using a different anchor modality by swapping the anchor and the contrasted modalities while keeping the weight design in Eqs. (20)–(21) unchanged.

D.1 Text-anchored form

If text is used as the anchor, the two InfoNCE terms become:

$$\ell_i^{T\to A} = -\log \frac{\exp\big(s(u_i^T, u_i^A)/\tau\big)}{\sum_{j=1}^{B} \omega_{ij} \exp\big(s(u_i^T, u_j^A)/\tau\big)}, \quad (56)$$

$$\ell_i^{T\to V} = -\log \frac{\exp\big(s(u_i^T, u_i^V)/\tau\big)}{\sum_{j=1}^{B} \omega_{ij} \exp\big(s(u_i^T, u_j^V)/\tau\big)}. \quad (57)$$

The gated batch objective follows Eq. (24) by replacing the two terms accordingly.

E Dataset Details

We conduct experiments on two multimodal datasets, T-MED and CMU-MOSEI.

T-MED is a large scale multimodal teacher sentiment analysis dataset with 14,938 instances. It contains eight teacher sentiment categories: neutral, anger, joy, surprise, sadness, patience, enthusiasm, expectation. We follow the label definitions and statistics reported in the original T-MED paper. The dataset is highly imbalanced, with neutral as the majority class. Specifically, neutral accounts for 7,318 samples (49.0%), followed by expectation with 2,493 samples (16.7%), joy with 1,619 samples (10.8%), patience with 916 samples (6.1%), enthusiasm with 834 samples (5.6%), anger with 821 samples (5.5%), surprise with 507 samples (3.4%), and sadness with 430 samples (2.9%). Following the original experimental setup, T-MED uses an 80%/10%/10% split for training/validation/test. Given 14,938 total instances, this corresponds to 11,950 training samples, 1,494 validation samples, and 1,494 test samples.

CMU-MOSEI is a benchmark for multimodal sentiment analysis and sentiment intensity prediction. In this work, we use its sentiment annotation on a 7 point Likert scale from -3 to $+3$: -3 (highly negative), -2 (negative), -1 (weakly negative), 0 (neutral), $+1$ (weakly positive), $+2$ (positive), and $+3$ (highly positive). The original paper visualizes the sentiment label distribution using histograms and explicitly notes that the plot shows ratios rather than absolute counts; it does not provide exact per class counts in a table. Therefore, we do not report exact per class sample counts here. We follow the commonly used predetermined split in prior work, consisting of 16,326 / 1,871 / 4,659 samples for train/validation/test, respectively.

F Implementation Details

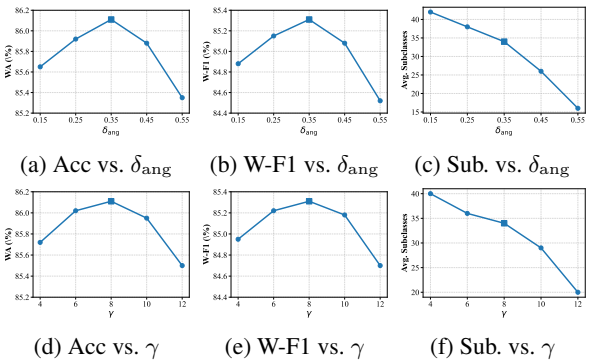
All experiments are run on a single NVIDIA RTX 4090 GPU with 24GB memory. We optimize with AdamW with a base learning rate of 2×10^{-4} and batch size 32. We train for up to 30 epochs with early stopping and patience 5. We repeat each experiment with five random seeds and report mean performance. Unless otherwise required by the original implementations, baselines follow the same feature extraction and preprocessing pipeline for fair comparison.

For the shared embedding space (Eq. (9)–(10)), we set the projection dimension to $d = 384$. For the CLS-guided cross-attention block (Eq. (12)–(13)), the multi-head attention uses 8 heads with dropout 0.1, and the FFN hidden dimension is set to 1536. For substyle discovery, we set the angular constraint threshold to $\delta_{ang} = 0.35$ and the BIC-gain threshold to $\gamma = 8.0$ for the thresholded acceptance step (Eq. (17)–(18)). For stability gating (Eq. (20)–(21)), we treat components with size smaller than $n_{min} = 20$ as tiny components. For

substyle-aware instance contrast (Eq. (23)–(28)), we set the temperature to $\tau = 0.07$ and use the four-level weights $\omega_{tiny} = 0.01$, $\omega_{sub} = 0.05$, $\omega_{cls} = 0.15$, and $\omega_{neg} = 1.0$, which satisfy $\omega_{tiny} < \omega_{sub} < \omega_{cls} < \omega_{neg}$ (Eq. (25)). For prototype contrast (Eq. (31)–(33)), we set the temperature to $\tau = 0.05$. Prototypes are updated with exponential moving average (EMA) (Eq. (36)) using momentum $\beta = 0.9$. For the angular margin softmax classifier (Eq. (38)), we set the scaling factor to $\alpha = 30$ and the angular margin to $m_{arc} = 0.20$. For the overall objective (Eq. (43)), we set the loss weights to $\lambda_{nce} = 0.08$, $\lambda_{proto} = 0.02$, and $\lambda_{sub} = 0.10$.

G Parameter Sensitivity Analysis

Substyle discovery uses two thresholds in the thresholded acceptance step: an angular constraint and a BIC-gain constraint $\Delta_c \geq \gamma$, where Δ_c measures the improvement of a multi-component GMM over a single-component fit. Figure 3 summarizes sensitivity. Across sweeps, the number of discovered substyles varies monotonically with the threshold, while ACC and W-F1 remain stable over a range. Loose settings (smaller δ_{ang} or γ) oversplit categories and degrade performance, whereas strict settings prune useful structure and weaken the benefit of modeling within-label substyles. Performance is consistently strong for $\delta_{\text{ang}} \in [0.25, 0.45]$ and $\gamma \in [6, 10]$.



(a) Acc vs. δ_{ang} (b) W-F1 vs. δ_{ang} (c) Sub. vs. δ_{ang}

(d) Acc vs. γ (e) W-F1 vs. γ (f) Sub. vs. γ

Figure 3: Sensitivity on T-MED ($T+A$) with one-factor-at-a-time sweeps: varying δ_{ang} (top row) with γ=8.0 fixed, and varying γ (bottom row) with δ_{ang}=0.35 fixed. We report ACC, W-F1, and the average number of discovered substyles over C=8 categories (Sub.).

H Baseline Details

General domain baselines. We include representative methods covering major paradigms in multimodal sentiment modeling. TFN and LMF perform tensor fusion and its low rank factorization. MFN augments fusion with a memory mechanism. MulT adopts cross modal Transformers to model inter modal interactions. MISA decomposes representations into modality invariant and modality specific subspaces. ConFEDE improves cross modal consistency via decomposed contrastive objectives. DFMU models uncertainty with distributional representations and distribution aware contrastive learning. DLF introduces language guided cross attention on top of disentangled representations. MFMB-Net improves robustness under complex conditions via a multi branch fusion design.

Teacher oriented baselines. To reflect classroom specific assumptions and data characteristics, we additionally include three teacher oriented methods. EDSN is a dual space network that learns a cross modal consistent space together with a modality discriminative space for teacher sentiment analysis. ST-SER targets smart classroom teacher speech sentiment analysis with multimodal cues. AAM-TSA emphasizes audio centric cues and teacher scene adaptation for teacher focused datasets.

Implementation and evaluation protocol. Unless otherwise specified by the original methods, all baselines are implemented using publicly released code or are faithfully reproduced from the authors' descriptions. All models are trained and evaluated in a unified experimental environment with consistent preprocessing and evaluation settings.

I Substyle and Prototype Statistics

We further report the scale of the discovered *stable* substyles and the corresponding prototype bank on T-MED under the $T+A$ setting. Substyles are counted after the stability gating and tiny-component filtering in Section 4.3. We maintain one prototype per class and one prototype per stable substyle following Section 4.4.

As shown in Table 4, the discovery procedure yields multiple recurrent substyles within each sentiment category (mean 4.25, max 6), suggesting substantial intra-class heterogeneity beyond a single coarse label. Meanwhile, the overall number of stable substyles remains moderate (34 in total), which keeps the prototype bank compact (8 class prototypes + 34 substyle prototypes) and thus computationally manageable.

Statistic	Value
Number of categories (C)	8
Substyles per category (min/mean/max)	2 / 4.25 / 6
Substyles per category (p10 / p50 / p90)	2.0 / 4.0 / 6.0
Total number of substyles	34
Substyle size (p10 / p50 / p90)	42 / 196 / 712
Number of prototypes (class / substyle)	8 / 34

Table 4: Statistics of discovered stable substyles and maintained prototypes on T-MED under $T+A$. Substyle sizes are reported after stability gating and tiny-component filtering as in Section 4.3.

J Additional Analysis for Grouped Ablations

This appendix complements the main text discussion by reporting additional difference based deltas from Table 2 that are not explicitly enumerated in the main text. All gains are in percentage points for ACC and W-F1.

Group 2 with the IC pathway. Under the same IC pathway, VAR-KM-IC improves over VAR-NULL by $+0.88$ ACC (84.45 vs. 83.57) and $+0.77$ W-F1 (83.21 vs. 82.44). Using DG with IC yields a larger margin over VAR-NULL. VAR-DG-IC gains $+1.47$ ACC (85.04 vs. 83.57) and $+1.87$ W-F1 (84.31 vs. 82.44).

Group 3 with the SC pathway. With SC fixed, DG provides consistent improvements over KM. VAR-DG-SC exceeds VAR-KM-SC by $+0.56$ ACC (84.79 vs. 84.23) and $+0.60$ W-F1 (83.72 vs. 83.12).

Group 4 with DG fixed. Comparing the two single constraint variants under DG, VAR-DG-IC is higher than VAR-DG-SC by $+0.25$ ACC (85.04 vs. 84.79) and $+0.59$ W-F1 (84.31 vs. 83.72). When adding SC on top of IC, VAR-DG-SCIC further improves over VAR-DG-IC by $+0.72$ ACC (85.76 vs. 85.04) and $+0.62$ W-F1 (84.93 vs. 84.31). Relative to VAR-NULL, the combined model VAR-DG-SCIC achieves an overall gain of $+2.19$ ACC (85.76 vs. 83.57) and $+2.49$ W-F1 (84.93 vs. 82.44).

Group 5 cumulative improvement. Beyond the incremental gain reported in the main text, the full configuration yields a net improvement of $+2.54$ ACC (86.11 vs. 83.57) and $+2.87$ W-F1 (85.31 vs. 82.44) over the minimal baseline VAR-NULL.

K Additional Case Analysis: Anger Pair

To further complement the quantitative results, this appendix provides a deeper qualitative analysis of the *Anger* pair (b1 vs. b2) in Figure 4. Although both instances share the same label, their differences map naturally onto our coarse-to-fine view of classroom sentiments: **coarse** granularity (performativity and regulated expression), **medium** granularity (within-label substyles), and **fine** granularity (context-dependent cross-modal cue compositions). We summarize how the same label can manifest regulated expression across modalities (performativity), within-label substyles, and context-dependent cross-modal cue compositions.

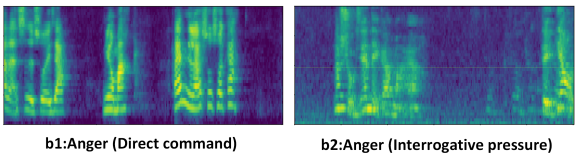


Figure 4: Case study: An additional example for the anger sentiment.

Case: Anger, b1 versus b2. Transcripts: (b1) *Go ahead and try to say it—do you have an answer? Come on, you tell us.* (b2) *So what should we do first? We need to investigate, right?*

Text reveals different pressure and control strategies. Instance (b1) resembles an imperative turn allocation that directly urges a response, while instance (b2) applies interrogative pressure to push an action plan. Audio provides complementary cues about how such pressure is enacted. The (b1) spectrogram is more punctuated, with stronger high-frequency components and sharper transitions, while (b2) is more sustained with energy concentrated in mid and low bands, consistent with pressing insistence rather than abrupt command.

At the *coarse* granularity under **performativity**, these two moments illustrate how anger in classrooms can be regulated and strategically expressed: teachers may keep certain channels relatively constrained while letting another channel carry the urgency. Here, the transcript makes the instructional pressure explicit, while audio differentiates the enacted style of that pressure (abrupt and punctuated vs. sustained and pressing), showing that regulated expression remains recoverable through selective cross-modal evidence.

At the *medium* granularity, **within-label heterogeneity** appears as reusable substyles within *Anger*. Instance (b1) reflects an imperative and commanding style, and instance (b2) reflects an interrogative and action-forcing style. Each substyle is associated with a different prosodic profile, with sharp

and punctuated delivery for (b1) and pressing and sustained delivery for (b2). Treating the label as a single center risks blurring such systematic within-class structure.

At the *fine* granularity, **contextual complementarity** is manifested as distinct yet repeatable cross-modal compositions under the same label. The two instances correspond to imperative $\times$ sharp/punctuated cues in (b1) and interrogative pressure $\times$ pressing/sustained cues in (b2). These patterns are informative and repeatable, indicating that cross-modal evidence is organized into structured and reusable cue compositions.

Together with the Expectation case analyzed in the main text, this Anger pair supports the three-property view of classroom sentiments. Evidence can be regulated yet recoverable through selective cross-modal retrieval, labels can be multi-centered with recurring substyles, and modalities can form stable, context-dependent compositions. These observations align with CF-TSA's goals of preserving unimodal structure while modeling cross-modal relations, discovering reliable within-label partitions, and consolidating recurring cue compositions.

L Additional Evidence on Performativity

This appendix provides supplementary evidence for the *performativity* assumption underlying our coarse-level formulation. Since the main text models teacher sentiment as a professionally regulated expression rather than a fully spontaneous affective outburst, it is important to further examine whether such regulation is not only theoretically plausible but also empirically observable in authentic classroom data.

L.1 Theoretical support

Performativity in classroom teaching is widely recognized in educational psychology as a professional characteristic shaped by instructional goals, role expectations, and display rules (Pekrun, 2006; Wang and Frenzel, 2025). Under this view, the observed classroom expression should not always be interpreted as a direct and unconstrained manifestation of internal affect; instead, it may reflect strategically regulated sentiment expression under pedagogical and professional constraints.

L.2 Annotation-based analysis

To further examine the empirical presence of performativity, we conducted a supplementary annotation study on T-MED. Specifically, we randomly sampled 100 instances from each of the eight sentiment categories, yielding 800 samples in total. We then asked three trained graduate students to independently judge whether each sample exhibited clear professional regulation in expression. All annotators have research backgrounds in AI for Education and related publication experience. If at least two annotators agreed that the teacher in a sample exhibited clear professional regulation, the sample was labeled as *performative*.

The annotation results show that performativity is not a marginal phenomenon in teacher scenarios, but a recurrent property of classroom sentiment expression, with an overall proportion of 32%. Across categories, the proportions are 53% for anger, 39% for joy, 37% for surprise, 34% for enthusiasm, 31% for sadness, 27% for patience, 22% for expectation, and 13% for neutral. These results suggest that professional regulation is broadly present in teacher sentiment expression, while its prevalence is unevenly distributed across different sentiment categories.

M Efficiency Analysis

This appendix complements the main empirical evaluation with an efficiency analysis of CF-TSA. Beyond predictive performance, practical multimodal systems for classroom scenarios should also exhibit a reasonable computational cost profile. We therefore report both training and inference efficiency to characterize the practical feasibility of the proposed framework under realistic deployment considerations.

M.1 Training efficiency

Although CF-TSA adopts a hierarchical learning strategy, its training pipeline remains computationally manageable. In particular, the model does not rely on iterative GMM optimization inside the gradient-based training loop. Instead, prototype representations are maintained through an Exponential Moving Average (EMA) mechanism, which provides an efficient online update strategy during training.

Benchmarks conducted on an NVIDIA RTX 4090 GPU (24GB) using the T-MED dataset under a unified tri-modal setting show that CF-TSA requires 153 seconds for training. This training cost is substantially lower than EDSN (212 seconds) and AAM-TSA (334 seconds), and remains close

to DLF (126 seconds), although it is higher than DMFU (108 seconds).

M.2 Inference efficiency

The inference overhead of CF-TSA remains lightweight because substyle discovery and contrastive optimization are restricted to the training stage. At inference time, the deployed architecture consists only of a CLS-guided cross-modal attention module and a classifier, without additional clustering or prototype optimization procedures.

On the 1,494-sample T-MED test set, CF-TSA requires 5.4 seconds in total, corresponding to 0.36 ms per sample. This inference cost is lower than MFMB-Net (6.2 seconds), EDSN (8.0 seconds), and AAM-TSA (12.5 seconds), and is close to DLF (5.3 seconds), although it remains slightly higher than DMFU (4.8 seconds). Overall, these results suggest that the proposed framework maintains a favorable trade-off between modeling capacity and computational cost, and is compatible with practical classroom deployment settings.