SPEAKER-DEPENDENT EMOTIONAL TRANSITION-AWARE MULTIMODAL DYNAMIC ANCHOR LEARNING FOR MULTIMODAL EMOTION RECOGNITION IN CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Emotion Recognition in Conversations (MERC) is a crucial task for improving a machine's emotional understanding. It aims to accurately infer emotions from the complex interplay of textual, acoustic, and visual signals in conversational context. Although existing methods have improved performance via advanced strategies like multimodal fusion, graph-based reasoning, etc., they often leave psychologically meaningful multimodal priors and speaker-aware emotion transitions out of consideration to guide multimodal fusion and alignment. In this paper, we propose a novel learning framework, **S**peaker-dependent **E**motional Transition-aware Multimodal **D**ynamic **A**nchor **L**earning (**SEDAL**), which introduces learnable class-level multimodal semantic anchors and speaker-aware emotional transition dynamics into the multimodal fusion and emotion recognition process. Specifically, SEDAL first constructs momentum-evolving multimodal emotion anchors, which serve as emotion-specific prototypes to guide the multimodal alignment among the text, audio, visual, and fused features. Then, SEDAL further extends these anchors from isolated utterances to adjacent conversational turns through speaker-aware emotion transition prototypes and applies anchor-guided modality perturbation consistency learning only on transition-relevant utterances. Finally, SEDAL further enhances recognition robustness with a deferred student-adaptive multi-teacher boundary distillation strategy, which transfers effective decisions from multiple validation-calibrated teachers. Extensive experiments on IEMOCAP and MELD datasets demonstrate that SEDAL achieves state-of-the-art performance, confirming its effectiveness for MERC.

1 INTRODUCTION

Multimodal Emotion Recognition in Conversations (MERC) aims to infer the underlying emotion expressed in each utterance of a dialogue by mining affective cues in textual, acoustic, and visual signals effectively. It plays a central role in various applications like affective conversation, human-computer interaction, social analysis, mental health estimation, and multimodal emotion understanding, but it is a challenging task since conversational emotion is a temporally evolving state shaped by speaker identity, inter-speaker influence, contextual dependency, and modality-specific uncertainty, rather than a static attribute of an isolated utterance. Therefore, an effective MERC model should accurately model current emotion cues in a multimodal utterance and the current emotion emergence in a multimodal conversational trajectory.

Early methods have made important progress in MERC by modeling dialogue context and speaker dependencies through recurrent speaker memories (Majumder et al., 2019), cognitive reasoning (Hu et al., 2021a), graph relations (Ghosal et al., 2019; Shen et al., 2021), Transformer interactions (Lian et al., 2021), and multimodal relational or dynamic fusion (Hu et al., 2021b; Joshi et al., 2022; Hu et al., 2022; Nguyen et al., 2023). However, they often simply treat emotion recognition as contextual feature aggregation followed by classification. To further improve recognition performance, recent studies have explored advanced technologies, like stronger modality balancing, adaptive fusion, graph/hypergraph-based reasoning, semantic anchoring, and distillation-based calibration. Modality

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.

balancing and self-distillation are used to improve the learning of weak modalities (Ma et al., 2024; Nguyen et al., 2024). Anchor-gated or low-rank matching attention mechanisms are deployed to reduce inefficient frame-level alignment and recalibrate cross-modal interaction (Li et al., 2025b; Shou et al., 2025). Graph, hypergraph, and global-local fusion are used to capture higher-order conversational relations(Wang et al., 2025; Shi et al., 2025; Meng et al., 2026; Li et al., 2025a). Prototype-based grounding introduces interpretable semantic references for multimodal emotion representations (Hu et al., 2025; Radford et al., 2021). Although these advances lift up MERC, existing methods either are fundamentally data-driven or only use visual anchors to guide the multimodal fusion and alignment process, leaving psychologically meaningful multimodal priors and speaker-aware emotion transitions out of consideration. This limits their robustness and generalization in real-world conversational settings where ambiguous, noisy, or weakly correlated multimodal inputs often occur. In addition, psychologically meaningful multimodal priors and speaker-aware emotion transitions should be modeled jointly. Otherwise, it is inaccurate for conversational emotion modeling, limiting a MERC model to learn a coherent geometry in which the current emotion, its conversational evolution, and its robustness under modality noise are mutually constrained. On the one hand, a multimodal prototype for a single utterance is a useful semantic anchor, but it does not by itself describe whether the current emotion is a continuation of the same speaker's prior state or a response to another speaker. On the other hand, an emotion transition model can capture temporal affective dynamics, but without a shared multimodal anchor space, it may learn context-sensitive patterns that are weakly tied to class-level emotion semantics.

In this paper, we propose **S**peaker-dependent **E**motional Transition-aware Multimodal **D**ynamic **A**nchor **L**earning (**SEDAL**), a unified learning framework that jointly models learnable class-level multimodal priors and speaker-aware emotional transition dynamics in multimodal emotion anchor space to guide the multimodal fusion and emotion recognition process. Specifically, to enhance modality-level semantic alignment, SEDAL introduces **Momentum-Evolving Multimodal Emotion Anchoring** (MEMEA) to project text, audio, visual, and fused representations into a shared anchor space and maintain class-level multimodal emotion anchors through exponential moving average updates. These class-level multimodal emotion anchors serve as emotion-specific prototypes to guide the multimodal alignment by making each modality representation and the fused representation close to their target emotion anchor while separating from other class anchors. Then, SEDAL introduces **Speaker-Aware Emotional Transition Prototype Learning with Perturbation Consistency Regularization** (STPL-PCR) to model the speaker-aware emotional transition dynamics by injecting speaker-aware adjacent-utterance transition features into multimodal emotion anchors to construct same-speaker and cross-speaker transition prototypes. STPL-PCR also applies **A**nchor-guided **M**odality **P**erturbation **C**onsistency (AMPC) regularization only to utterances with valid transition context to enhance generalization. To further enhance recognition robustness by refining decision boundaries without making inference depend on an ensemble, SEDAL introduces **Deferred Student-Adaptive Multi-Teacher Boundary Distillation** (DSA-MTBD) to transfer high-confidence soft distributions from multiple validation-calibrated teachers into a single student. Extensive experiments on IEMOCAP and MELD datasets demonstrate that the SEDAL achieves state-of-the-art performance, confirming its effectiveness for MERC.

Our contributions are as follows. First, we introduce SEDAL to jointly model learnable class-level multimodal priors and speaker-aware emotional transition dynamics in a multimodal emotion anchor space to guide the multimodal fusion and emotion recognition process. Second, we propose MEMEA to learn class-level multimodal prototypes and align different modalities in a common emotion-anchor space. Third, we propose STPL-PCR to extend utterance-level emotion anchors to speaker-aware transition prototypes. Fourth, we propose DSA-MTBD to transfer reliable multi-teacher decision-boundary information to a single student for further improving generalization without the cost of multi-model ensemble inference. Finally, extensive experiments show the SEDAL achieves state-of-the-art performance on multimodal emotion recognition.

2 Related Work

Multimodal Emotion Recognition in Conversations. MERC interprets each utterance through speaker identity, dialogue context, and multimodal evidence. Early methods model speaker and temporal dependencies with recurrent memory or utterance graphs (Majumder et al., 2019; Ghosal et al., 2019; Shen et al., 2021). Multimodal methods extend this foundation through cross-modal

interaction, graph propagation, speaker-aware context, and dynamic fusion (Lian et al., 2021; Hu et al., 2021b; Joshi et al., 2022; Hu et al., 2022). Recent work refines temporal reasoning, modality weighting, graph fusion, distillation, and sentiment dynamics (Nguyen et al., 2023; Ma et al., 2024; Nguyen et al., 2024; Wang et al., 2025; Shi et al., 2025; Li et al., 2025b; Meng et al., 2026; Li et al., 2025a; 2026). These approaches primarily supervise the final emotion prediction without an explicit class-level semantic target shared across modality branches. SEDAL instead uses a shared multimodal anchor geometry to connect utterance-level emotion semantics with speaker-aware transition prototypes and guide robustness through emotional transitions.

Multimodal Fusion and Alignment. Multimodal fusion seeks shared semantics while retaining modality-specific cues. MulT learns cross-modal attention for unaligned sequences (Tsai et al., 2019), while MISA, Self-MM, and DMD use factor separation, unimodal supervision, or decoupled distillation to preserve distinct modality information (Hazarika et al., 2020; Yu et al., 2021; Li et al., 2023). Recent methods introduce transferable visual-semantic anchors through CLIP and VEGA (Radford et al., 2021; Hu et al., 2025), efficient cross-modal attention (Shou et al., 2025), hierarchical alignment of common and unique representations (Qian et al., 2026), and cross-modal training synergy under weak modality pairing (Lee and Yoon, 2025). Distillation transfers predictions or modality knowledge across models (Hinton et al., 2015; Li et al., 2023; 2025b). These approaches mainly align utterance-level representations or logits. SEDAL links task-specific multimodal emotion anchors to speaker-aware transition prototypes and applies selective distillation after the student anchor space stabilizes.

3 METHOD

SEDAL aims to jointly model learnable class-level multimodal priors and speaker-aware emotional transition dynamics in multimodal emotion anchor space to guide the multimodal fusion and emotion recognition process. Thus, SEDAL builds a task-specific multimodal emotion-anchor space and grounds the multimodal fusion and emotion recognition of the MERC backbone with class-level multimodal semantic anchors and speaker-aware emotional transition dynamics by three key modules during training: **Momentum-Evolving Multimodal Emotion Anchoring** (MEMEA), **Speaker-Aware Emotional Transition Prototype Learning with Perturbation Consistency Regularization** (STPL-PCR), and **Deferred Student-Adaptive Multi-Teacher Boundary Distillation** (DSA-MTBD). The MERC backbone produces context-aware unimodal representations and multimodal representations. MEMEA projects unimodal representations and multimodal representations into a shared anchor space and maintains class-level multimodal emotion anchors with exponential moving average updates. STPL-PCR models speaker-aware emotional transition dynamics by constructing speaker-aware adjacent-utterance transition representations based on the MEMEA anchor features and constructs same-speaker and cross-speaker transition prototypes. STPL-PCR also applies AMPC only to utterances with valid transition context to enhance generalization. DSA-MTBD transfers high-confidence soft distributions from multiple validation-calibrated teachers into a single student to further enhance recognition robustness by refining decision boundaries without making inference depend on an ensemble.

3.1 MERC BACKBONE

Let $\mathcal{D} = \{(\mathbf{u}_t, s_t, y_t)\}_{t=1}^{N}$ is a dialogue with N utterances, where u_t is the t-th utterance, s_t is its speaker identity, and y_t is the emotion label. Each utterance u_t is defined as a tuple of modality-specific inputs: $u_t = (\mathbf{x}_t^T, \mathbf{x}_t^A, \mathbf{x}_t^V)$, where $\mathbf{x}_t^T$, $\mathbf{x}_t^A$, and $\mathbf{x}_t^V$ represent the raw textual, acoustic, and visual signals of the t-th utterance, respectively. Following standard MERC pipelines (Ma et al., 2024; Hu et al., 2025), each modality-specific input is first passed through a dedicated encoder followed by a Conv1D with kernel 1 to extract utterance-level unimodal features in a shared hidden dimension d_h:

$$\mathbf{e}_t^m = \mathrm{Conv1D}_m(F_{enc}^m(\mathbf{x}_t^m)), \quad m \in \mathcal{M} = \{T, A, V\} \tag{1}$$

where F_{enc}^T, F_{enc}^A, and F_{enc}^V refer to RoBERTa, OpenSMILE, and DenseNet-based encoders, respectively. For the whole dialogue, the unimodal representation of utterance of modality m is defined as $\mathbf{E}^m = [\mathbf{e}_1^m, \ldots, \mathbf{e}_N^m]$. Then, a modality-specific contextual encoder with 9 stacked Cross-Modal Transformers and a Unimodal Gated Fusion is applied to obtain contextual representations:

$$\mathbf{H}^m = \mathrm{ContextualEncoder}_m(\mathbf{E}^m), \quad \mathbf{H}^m = [\mathbf{h}_1^m, \ldots, \mathbf{h}_N^m], \tag{2}$$

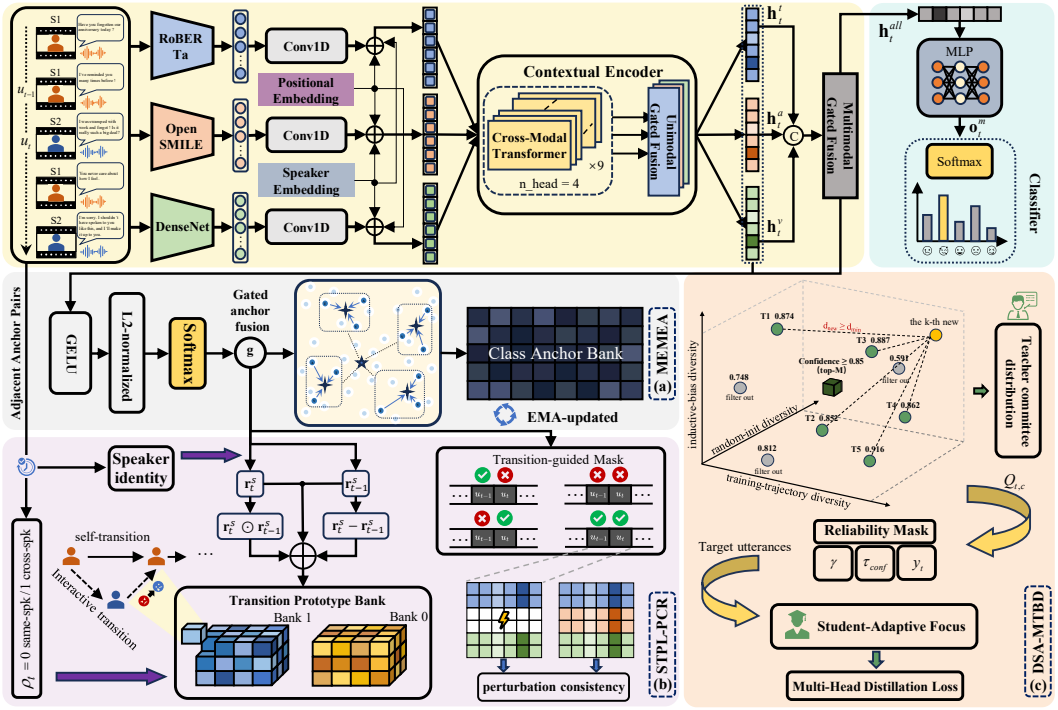


Figure 1: Overall architecture of the proposed SEDAL. Except for the MERC backbone, SEDAL consists of three main components: (a) MEMEA learns momentum-evolving multimodal emotion anchors for class-aware alignment, (b) STPL-PCR models same- and cross-speaker emotional transition prototypes with perturbation consistency, and (c) DSA-MTBD distills reliable multi-teacher boundary knowledge into a single student. The final multimodal representation is used for emotion classification.

where $\mathbf{h}_t^m \in \mathbb{R}^{d_h}$ is the contextual representation of utterance u_t under modality m. The multimodal representation $\mathbf{h}_t^M$ is obtained through a Multimodal Gated Fusion:

$$\mathbf{h}_t^M = \sum_{m \in \mathcal{M}} \pi_t^m \mathbf{h}_t^m, \quad \boldsymbol{\pi}_t = \mathrm{softmax}\left(W_f[\mathbf{h}_t^T; \mathbf{h}_t^A; \mathbf{h}_t^V] + \mathbf{b}_f\right), \tag{3}$$

where $[\cdot;\cdot]$ denotes concatenation, $\boldsymbol{\pi}_t$ is the modality-gating vector for utterance u_t, and W_f and $\mathbf{b}_f$ are learnable fusion parameters. For each representation in $\mathbf{h}_t^m$ ($m \in \mathcal{M}+ = \{T, A, V, M\}$), we can produce its emotion logits $\mathbf{o}_t^m \in \mathbb{R}^C$ and probability distribution $\mathbf{p}_t^m \in \mathbb{R}^C$ over C emotion classes:

$$\mathbf{o}_t^m = W_o^m \mathbf{h}_t^m + \mathbf{b}_o^m, \quad \mathbf{p}_t^m = \mathrm{softmax}(\mathbf{o}_t^m). \tag{4}$$

The final emotion prediction of utterance u_t is produced only by $\mathbf{p}_t^M$, while the modality-specific probabilities $\mathbf{p}_t^m$ ($m \in T, A, V$) will be used in the DSA-MTBD module. Let $\mathcal{U}$ be the set of valid utterances in the mini-batch and CE be the cross-entropy loss, a classification loss $\mathcal{L}_{cls}$ is depolyed:

$$\mathcal{L}_{cls} = \frac{1}{|\mathcal{U}|} \sum_{(i,t) \in \mathcal{U}} \mathrm{CE}(\mathbf{o}_{i,t}^M, y_{i,t}), \tag{5}$$

3.2 MEMEA

Momentum-Evolving Multimodal Emotion Anchoring (MEMEA) aims to construct multimodal emotion anchors, which serve as emotion-specific prototypes to guide the multimodal alignment among the text, audio, visual, and multimodal features. For each utterance t, contextual representation $\mathbf{h}_t^m$ ($m \in \mathcal{M}+$) is projected to the anchor space and normalized as $\mathbf{z}_t^m = \frac{f_m(\mathbf{h}_t^m)}{\|f_m(\mathbf{h}_t^m)\|_2} \in \mathbb{R}^{d_z}$, where $f_m(\cdot)$ is the projection head of modality m. Then, the utterance-level multimodal anchor feature $\mathbf{r}_t \in \mathbb{R}^{d_z}$ is obtained by a gated fusion over unimodal features $\mathbf{z}_t^m$ ($m \in \mathcal{M}$):

$$\boldsymbol{\alpha}_t = \mathrm{softmax}(g([\mathbf{z}_t^t; \mathbf{z}_t^a; \mathbf{z}_t^v])), \quad \mathbf{r}_t = \Big(\sum_{m \in \mathcal{M}} \alpha_t^m \mathbf{z}_t^m\Big) \Big/ \Big\| \sum_{m \in \mathcal{M}} \alpha_t^m \mathbf{z}_t^m \Big\|_2, \tag{6}$$

where $g(\cdot)$ is a gating network and $\boldsymbol{\alpha}_t$ is the modality weight vector in the anchor space. The multimodal anchor feature is used to update the class anchor bank $\mathbf{P} \in \mathbb{R}^{C\times d_z}$, where $\mathbf{p}_c$ is the prototype of emotion class c. Accurate class prototypes are crucial for grounding multimodal fusion and alignment, so we use learnable prototypes updated by an exponential moving average, rather than static prototypes:

$$\mathbf{p}_c \leftarrow \frac{\mu_{ma}\mathbf{p}_c + (1-\mu_{ma})\bar{\mathbf{r}}_c}{\|\mu_{ma}\mathbf{p}_c + (1-\mu_{ma})\bar{\mathbf{r}}_c\|_2}, \tag{7}$$

where $\mu_{ma} \in [0, 1)$ is the momentum coefficient and $\bar{\mathbf{r}}_c$ is class-wise anchor center of a mini-batch $\mathcal{B}$, which can be computed as $\bar{\mathbf{r}}_c = (\sum_{(i,t)\in\mathcal{B}} \mathbb{I}[y_{i,t} = c]\mathbf{r}_{i,t})/(\sum_{(i,t)\in\mathcal{B}} \mathbb{I}[y_{i,t} = c] + \epsilon)$. i, t denotes the t-th utterance of the i-th dialogue in the mini-batch, $c \in \mathcal{C} = \{1, \ldots, C\}$ is an emotion class, $\mathbb{I}[\cdot]$ is the indicator function, and ϵ avoids division by zero. The $\mathbf{p}_c$ will be updated if class c appears in the mini-batch. When class c first appears during training, $\mathbf{p}_c$ is set to $= \bar{\mathbf{r}}_c / \|\bar{\mathbf{r}}_c\|_2$. To optimize the prototypes in the class anchor bank $\mathbf{P}$ to make the anchor space a stable semantic substrate for emotion transition semantic injection, for each $m \in \mathcal{M}+$, we first compute class-level prototype logits as temperature- scaled value between $\mathbf{z}_t^m$ and each class anchor $\mathbf{p}_c$: $\ell_{t,c}^m = \tau_{ma}^{-1} \frac{(\mathbf{z}_t^m)^\top \mathbf{p}_c}{\|\mathbf{p}_c\|_2}$, where τ_{ma} is the anchor temperature. Then, a MEMEA loss $\mathcal{L}_{ma}$, which is a cross-entropy-based supervised prototype classification objective, is deployed to optimize the class prototypes. MEMEA applies supervised prototype classification loss $\mathcal{L}_{ma}$ to all prototype logits with modality weights ω_m:

$$\mathcal{L}_{ma} = \sum_{m\in\mathcal{M}^+} \omega_m \frac{1}{|\mathcal{U}|} \sum_{(i,t)\in\mathcal{U}} \mathrm{CE}(\boldsymbol{\ell}_{i,t}^m, y_{i,t}), \tag{8}$$

3.3 STPL-PCR

Speaker-Aware Emotional Transition Prototype Learning with Perturbation Consistency Regularization (STPL-PCR) aims to inject speaker-aware emotional transition dynamics into utterance-level multimodal emotion anchors. It learns speaker-aware transition prototypes from adjacent turns and applies AMPC to transition-related utterances to improve their robustness, where contextual ambiguity and modality conflict are more likely to affect recognition. For each valid adjacent pair $(\mathbf{u}_{t-1}, \mathbf{u}_t)$, STPL-PCR first augments the MEMEA anchor with a speaker embedding, $\mathbf{r}_t^s = [\mathbf{r}_t; \mathbf{e}_{s_t}]$, and constructs the transition input and normalized transition representation $\mathbf{d}_t$ as

$$\mathbf{d}_t = f_{tr}(\mathbf{q}_t)/\|f_{tr}(\mathbf{q}_t)\|_2, \quad \mathbf{q}_t = [\mathbf{r}_t^s; \mathbf{r}_{t-1}^s; \Delta\mathbf{r}_t^s; \mathbf{r}_t^s \odot \mathbf{r}_{t-1}^s]. \tag{9}$$

Here $\Delta\mathbf{r}_t^s = \mathbf{r}_t^s - \mathbf{r}_{t-1}^s$, $\odot$ denotes element-wise multiplication, and $f_{tr}(\cdot)$ is a single linear projection (followed by ℓ_2 normalization) that maps the adjacent-pair descriptor into the transition-anchor space.

Like MEMEA, STPL-PCR maintains a speaker-conditioned prototype bank $\mathbf{A} \in \mathbb{R}^{2\times C\times d_{tr}}$, where $\mathbf{a}_{\rho,c}$ represents emotion class c under speaker relation ρ. The $\mathbf{a}_{\rho,c}$ is also updated by EMA:

$$\mathbf{a}_{\rho,c} \leftarrow (\mu_{tr}\mathbf{a}_{\rho,c} + (1-\mu_{tr})\bar{\mathbf{d}}_{\rho,c})/\left\|\mu_{tr}\mathbf{a}_{\rho,c} + (1-\mu_{tr})\bar{\mathbf{d}}_{\rho,c}\right\|_2, \tag{10}$$

where $\mu_{tr} \in [0, 1)$ is the transition-prototype momentum and $\bar{\mathbf{d}}_{\rho,c}$ is speark-aware class-specific transition center of a mini-batch transition set $\mathcal{T}$. The $\bar{\mathbf{d}}_{\rho,c}$ is computed as $(\sum_{(i,t)\in\mathcal{T}} \mathbb{I}[\rho_{i,t} = \rho]\mathbb{I}[y_{i,t} = c]\mathbf{d}_{i,t})/(\sum_{(i,t)\in\mathcal{T}} \mathbb{I}[\rho_{i,t} = \rho]\mathbb{I}[y_{i,t} = c] + \epsilon)$ and $\mathcal{T} = \{(i, t) : t \geq 2,\ \delta_{i,t-1} = 1,\ \delta_{i,t} = 1\}$, where $\delta_{i,t} \in \{0, 1\}$ indicate whether utterance u_t in the i-th dialogue of a mini-batch is valid. This mean that $\mathcal{T}$ excludes the first utterance and any padding position. For each $(i, t) \in \mathcal{T}$, the speaker relation is $\rho_t = 0$ for same-speaker transitions, that is, speaker identify $s_{i,t}$ is same as $s_{i,t-1}$. Otherwise, the speaker relation is $\rho_t = 1$ for cross-speaker transition. Given the observed relation, the transition logits $\eta_{t,c}$ are computed as $\tau_{tr}^{-1}\mathbf{d}_t^\top \mathbf{a}_{\rho_t,c}$, where τ_{tr} is the transition temperature. To optimize the transition prototypes, STPL-PCR uses $\mathcal{L}_{tr}$:

$$\mathcal{L}_{tr} = \frac{1}{|\mathcal{T}|} \sum_{(i,t)\in\mathcal{T}} \mathrm{CE}(\boldsymbol{\eta}_{i,t}, y_{i,t}), \tag{11}$$

To improve the robustness of transition prototypes, we apply AMPC to perturb only utterances that have a valid preceding turn and enforce the prediction based on perturbed features and the prediction

based on original features be consistent. For a selected transition candidate, modality $\kappa_t \in \mathcal{M}$ is randomly masked, and the perturbed modality-specific representations $\tilde{\mathbf{h}}_t^m$ are passed through the same fusion layer and MEMEA heads to obtain $\tilde{\mathbf{r}}_t$, $\tilde{\boldsymbol{\ell}}_t^m$, and $\tilde{\mathbf{o}}_t^{all}$. The computation procedures of $\tilde{\mathbf{r}}_t$, $\tilde{\boldsymbol{\ell}}_t^m$, and $\tilde{\mathbf{o}}_t^{all}$ are the same with $\mathbf{r}_t$, $\boldsymbol{\ell}_t^m$, and $\mathbf{o}_t^{all}$. Then, we deploy a consistency loss to further optimize transition prototypes. Let $\mathcal{P}$ be the set of actually perturbed utterances, the full AMPC objective is

$$\mathcal{L}_{pc} = \beta_{anc}\mathcal{L}_{pc}^{anc} + \beta_{proto}\mathcal{L}_{pc}^{proto} + \beta_{cls}\mathcal{L}_{pc}^{cls}, \tag{12}$$

where β_{anc}, β_{proto}, and β_{cls} weight the anchor-feature consistency term $\mathcal{L}_{pc}^{anc}$, the prototype-logit consistency term $\mathcal{L}_{pc}^{proto}$, and the classifier-logit consistency term $\mathcal{L}_{pc}^{cls}$. The $\mathcal{L}_{pc}^{anc}$, $\mathcal{L}_{pc}^{proto}$, and $\mathcal{L}_{pc}^{cls}$ are computed as follows (sg denotes stop gradients):

$$\mathcal{L}_{pc}^{anc} = \frac{1}{|\mathcal{P}|}\sum_{(i,t)\in\mathcal{P}}\left(1 - \tilde{\mathbf{r}}_{i,t}^{\top}\mathrm{sg}(\mathbf{r}_{i,t})\right), \tag{13}$$

$$\mathcal{L}_{pc}^{proto} = \frac{1}{|\mathcal{P}|}\sum_{(i,t)\in\mathcal{P}}\mathrm{KL}\left(\mathrm{softmax}(\mathrm{sg}(\boldsymbol{\ell}_{i,t}^{anc})/\tau_{pc}) \,\middle\|\, \mathrm{softmax}(\tilde{\boldsymbol{\ell}}_{i,t}^{anc}/\tau_{pc})\right), \tag{14}$$

$$\mathcal{L}_{pc}^{cls} = \frac{1}{|\mathcal{P}|}\sum_{(i,t)\in\mathcal{P}}\mathrm{KL}\left(\mathrm{softmax}(\mathrm{sg}(\mathbf{o}_{i,t}^{all})/\tau_{pc}) \,\middle\|\, \mathrm{softmax}(\tilde{\mathbf{o}}_{i,t}^{all}/\tau_{pc})\right). \tag{15}$$

3.4 DSA-MTBD

Deferred Student-Adaptive Multi-Teacher Boundary Distillation (DSA-MTBD) aims to transfer reliable decision-boundary information from a set of validation-calibrated teachers into a single student. It treats the teachers as a training-only boundary oracle, filters noisy teacher targets, and delays distillation until the anchor and transition geometry has become stable. Let $\mathcal{K}$ be the number of teachers, the teacher distribution $\mathbf{Q}_t$ for utterance $\mathbf{u}_t$ is: $\mathbf{Q}_t = \sum_{k\in\mathcal{K}}\alpha_k\sum_{m\in\mathcal{M}+}\beta_m\mathbf{p}_{k,t}^m$, where $\sum_{k\in\mathcal{K}}\alpha_k = 1$, $\sum_{m\in\mathcal{M}+}\beta_m = 1$, $\mathbf{p}_{k,m}(\mathbf{u}_t)$ is the probability vector predicted by teacher k and modality m. The teacher pool is diversified along training trajectory, random initialization, and objective-level inductive bias. To avoid admitting near-duplicate teachers, the marginal diversity d_{new} of a new candidate k_{new} is computed on multimodal-fused validation predictions: $d_{new} = \frac{1}{|\mathcal{K}_{pool}|}\sum_{j\in\mathcal{K}_{pool}}\frac{1}{|\mathcal{V}|}\sum_{u_t\in\mathcal{V}}\mathrm{JS}(\mathbf{p}_{new,\mathrm{M}}(u_t), \mathbf{p}_{j,\mathrm{M}}(u_t))$. Candidates that pass this diversity gate are then weighted by validation accuracy. All teachers are frozen once they are trained. DSA-MTBD retains $\mathbf{Q}_t$ only when its top class matches y_t, and its confidence exceeds $\tau_{conf} = 0.85$. The retained teacher distribution is weighted by $v_t = (\max_c Q_{t,c})^{\gamma}$, where $\gamma = 0.8$. It also applies a student-adaptive focus weight $r_t = (1 - p_s^M(\hat{c}_t \mid u_t))^{\rho}$, where $\hat{c}_t = \arg\max_c Q_{t,c}$, so distillation concentrates on samples whose teacher boundary the current all-modal student has not yet matched. The teacher distribution is mildly smoothed as $\bar{\mathbf{Q}}_t = (1-\varepsilon)\mathbf{Q}_t + \varepsilon/C$, where ε is the smoothing coefficient. The distillation starts after epoch e_0 with a warmup over w epochs and the distillation coefficient $\lambda_{kd}(e)$ is defined as $\lambda_{kd}(e) = \lambda_{kd}^{max}\min\left(1, \max\left(0, \frac{e-e_0+1}{w}\right)\right)$, where e is the current epoch. DSA-MTBD distills the teacher target into a student model as follows:

$$\mathcal{L}_{kd} = \frac{1}{|\mathcal{U}|}\sum_{(i,t)\in\mathcal{U}} v_{i,t}r_{i,t}\sum_{m\in\mathcal{M}^+}\mathrm{KL}\left(\bar{\mathbf{Q}}_{i,t} \,\middle\|\, \mathrm{softmax}(\mathbf{o}_{i,t}^m/T_{kd})\right), \tag{16}$$

3.5 Training Objective

SEDAL is optimized by combining the backbone classification objective with the three proposed modules. Thus, the overall objective is

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{ma}\mathcal{L}_{ma} + \lambda_{tr}\mathcal{L}_{tr} + \lambda_{pc}\mathcal{L}_{pc} + \lambda_{kd}(e)\mathcal{L}_{kd}, \tag{17}$$

where e is the training epoch and the $\lambda(e)$ terms schedule the module weights. MEMEA is warmed up first, STPL-PCR and AMPC are introduced after anchor features become available, and DSA-MTBD starts only after the deferred epoch so that teacher boundary supervision refines an already formed anchor geometry.

Table 1: Performance (%) on IEMOCAP (left) and MELD (right). We report F1 for each emotion. Acc and W-F1 denote the overall accuracy and weighted F1, respectively. "–" marks values not reported in the original paper. **Best** results are highlighted in **bold**, and second-best results are highlighted underlined. For SEDAL, we report the mean over 5 runs, with the standard deviation printed below each value. † denotes reproduced results.

Methods	IEMOCAP								MELD								
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	W-F1	Joy	Sadness	Neutral	Anger	Surprise	Disgust	Fear	Acc	W-F1
SDT†	64.24	81.36	74.35	68.35	80.63	68.33	74.03	74.12	64.38	**43.90**	80.14	54.17	58.64	27.27	18.92	67.55	66.64
D^2GCN†	60.51	80.63	68.12	66.41	73.67	66.71	70.80	70.09	57.36	28.19	76.57	46.80	50.73	2.30	0.00	61.65	59.81
HAUCL†	53.48	81.71	69.59	65.50	76.06	68.16	70.84	70.91	64.14	39.01	79.85	55.18	59.36	27.78	17.28	67.43	66.15
AdaIGN	53.04	81.47	71.26	65.87	76.34	67.79	70.49	70.74	64.54	43.70	79.75	56.15	60.53	–	–	67.62	66.79
CMERC†	56.87	80.92	70.31	68.35	78.37	65.80	70.93	71.13	64.23	38.97	79.85	54.21	61.10	30.16	21.69	67.32	66.59
DER-GCN	58.80	79.80	61.50	**72.10**	73.30	67.80	69.70	69.40	–	–	–	–	–	–	–	66.80	66.10
MAGTKD†	56.92	80.83	67.60	68.38	76.32	67.69	69.85	69.96	61.48	40.44	79.18	52.20	56.95	27.18	20.90	66.09	64.94
ECERC†	60.25	78.28	72.57	66.46	79.08	67.76	71.60	71.77	64.28	40.83	79.94	53.62	58.09	30.07	**26.27**	67.55	66.58
MATCH	59.18	82.10	74.44	67.28	79.67	66.31	72.64	72.55	63.75	41.82	80.38	**56.98**	**61.34**	26.67	19.75	68.16	67.04
GraphSmile†	62.40	82.69	70.18	69.85	78.44	67.03	72.14	72.29	63.57	42.82	79.92	53.13	58.72	27.72	16.90	67.20	66.06
BIG-FUSION†	62.49	80.75	73.10	68.72	76.30	68.60	71.96	72.62	62.96	36.48	80.09	51.41	56.99	0.00	0.00	64.18	63.79
DIB-HGCN	–	–	–	–	–	–	72.58	72.46	–	–	–	–	–	–	–	68.01	66.61
VEGA†	**67.87**	80.98	73.43	68.61	79.18	70.30	73.77	73.90	64.05	41.01	81.16	52.23	57.65	27.61	24.42	65.45	66.49
DF-GCN†	62.04	80.87	70.79	69.81	75.65	68.81	73.12	72.08	63.96	42.63	80.38	54.45	58.80	24.98	20.39	66.90	66.84
CMDCD†	58.82	79.76	73.15	68.89	77.51	70.25	72.53	72.70	61.80	38.54	80.95	52.68	57.25	26.45	21.51	67.02	65.79
AO-FL†	62.94	80.37	72.82	67.95	78.41	68.73	71.87	72.44	63.96	42.90	79.76	53.89	57.34	27.07	21.97	66.90	66.28
DnR†	62.16	78.59	75.57	68.20	78.58	66.65	72.17	72.53	59.30	39.80	79.61	51.18	51.65	16.00	14.51	65.83	64.08
BiSSM-CL	55.17	**83.20**	74.78	68.53	80.90	69.97	73.57	73.66	64.35	40.81	79.69	54.27	58.67	–	–	67.21	65.94
SEDAL (Ours)	64.45 ±1.93	81.74 ±0.45	**75.85** ±0.48	69.97 ±2.21	**81.03** ±0.95	**71.67** ±1.81	**74.98** ±0.26	**75.09** ±0.32	**64.76** ±1.03	43.42 ±1.70	**81.69** ±0.62	56.82 ±1.34	59.79 ±1.55	**30.68** ±2.47	25.01 ±3.13	**68.74** ±0.29	**67.98** ±0.23

4 Experiments and Analysis

4.1 Experimental Settings

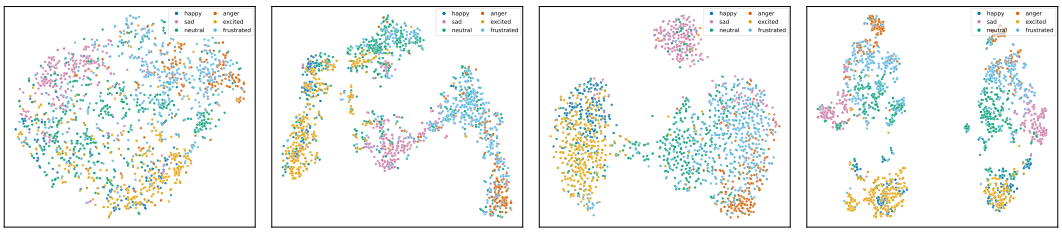
We evaluate SEDAL on two widely adopted MERC benchmarks, IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019), both of which provide synchronized textual, acoustic, and visual information for conversational emotion analysis. Following the common evaluation protocol in MERC, we report the overall accuracy (Acc) and the F1-score of each emotion class, together with the weighted average F1-score (W-F1), so that the comparison reflects both aggregate performance and class-wise recognition quality. We compare SEDAL with 18 recent and representative MERC baselines, including SDT (Ma et al., 2024), D^2GCN (Dai et al., 2024), DER-GCN (Ai et al., 2025), AdaIGN (Tu et al., 2024a), GraphSmile (Li et al., 2025a), DF-GCN (Meng et al., 2026), HAUCL (Yi et al., 2024), DIB-HGCN (Chen and Shi, 2025), MATCH (Shi et al., 2025), CMERC (Tu et al., 2024b), ECERC (Zhang and Tan, 2025), MAGTKD (Li et al., 2025b), VEGA (Hu et al., 2025), BIG-FUSION (Wang et al., 2025), BiSSM-CL (Li et al., 2026), CMDCD (Zhu et al., 2026), AO-FL (Che et al., 2026), and DnR (Mai et al., 2026). More details about datasets, metrics, baselines, and implementation details can refer to Appendix C, Appendix D, Appendix E, and Appendix F.

4.2 Results and Analysis

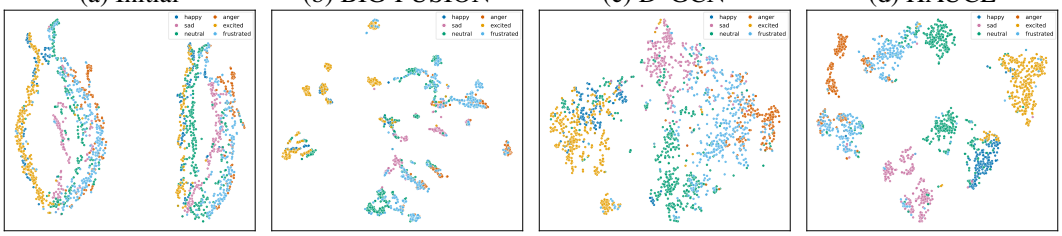
As shown in Table 1, SEDAL achieves the best overall performance on both benchmarks, with 74.98 Acc and 75.09 W-F1 on IEMOCAP and 68.74 Acc and 67.98 W-F1 on MELD. It improves W-F1 over the strongest baselines by 0.97 and 0.94, respectively, while also attaining the highest accuracy on both datasets. These gains over graph-based, hypergraph-based, distillation-based, and fusion-oriented methods suggest that explicitly modeling multimodal emotional dynamics is beneficial beyond backbone strength. At the class level, SEDAL ranks first or second for most emotions. It performs particularly well on Neutral, Excited, and Frustrated in IEMOCAP while remaining competitive on Happy and Angry, suggesting effectiveness across stable emotional states and speaker-dependent shifts. On MELD, it achieves the best scores on Joy, Neutral, and Disgust and the second-best scores on Anger and Fear. These balanced improvements are consistent with the complementary roles of class-aware multimodal anchors, speaker-aware transition prototypes, and perturbation consistency under incomplete or noisy modality evidence.

4.3 Ablation Studies

Table 2 presents the component- and modality-wise ablations. Removing any proposed module degrades performance on both benchmarks, showing that SEDAL's gains arise from



(a) Initial (b) BIG-FUSION (c) D^{2}GCN (d) HAUCL



(e) CMDCD (f) CMERC (g) ECERC (h) SEDAL

Figure 2: t-SNE visualization of utterance-level features for MERC on the IEMOCAP dataset. We visualize the feature distributions learned by raw input. Each point represents one utterance, and different colors indicate distinct emotion categories.

their complementary contributions. Removing STPL-PCR ($\mathcal{L}_{tr} + \mathcal{L}_{pc}$) causes the largest W-F1 drop, from 75.09 to 72.69 on IEMOCAP and from 67.98 to 65.47 on MELD. This result supports explicitly modeling same- and cross-speaker emotional transitions with transition-guided perturbation consistency for context-dependent and modality-ambiguous utterances. Removing MEMEA ($\mathcal{L}_{ma}$) also causes a clear decline, particularly on IEMOCAP, demonstrating the value of a stable class-level geometry for aligning unimodal and fused representations and guiding the transition and robustness objectives. Removing DSA-MTBD ($\mathcal{L}_{kd}$) produces a smaller but consistent decline, indicating that deferred multi-teacher boundary distillation further calibrates the student after the anchor geometry forms while adding no inference overhead. The modality ablations reveal a text-dominant pattern: text is the strongest single modality, whereas audio and visual cues alone are substantially weaker, especially on MELD. Nevertheless, combining these nonverbal cues with text improves performance, and the full trimodal model performs best, demonstrating their complementary value within the multimodal anchor and speaker-aware transition framework.

Table 2: Ablation results (%) on IEMOCAP and MELD.

	Configuration	IEMOCAP		MELD	
		Acc	W-F1	Acc	W-F1
	SEDAL	**74.98**	**75.09**	**68.74**	**67.98**
Component	w/o MEMEA	72.79	73.35	67.09	66.53
	w/o STPL-PCR	72.06	72.69	66.12	65.47
	w/o DSA-MTBD	73.81	74.32	68.29	67.32
Modality	Text	68.87	69.23	66.74	65.79
	Audio	63.45	64.07	50.56	43.25
	Visual	44.93	45.48	49.43	36.02
	Text + Audio	73.12	73.24	67.31	66.94
	Text + Visual	70.51	70.76	67.36	66.72
	Audio + Visual	64.27	64.59	50.24	42.36

4.4 Multi-Modal Representation Visualization

To examine whether the quantitative gains correspond to a more discriminative representation geometry, we visualize the utterance-level fused features on the IEMOCAP test set using t-SNE. As shown in Figure 2, the initial features exhibit substantial label interleaving: samples form broad manifolds, but their local neighborhoods are not consistently emotion-aware. Although the baselines improve this structure to varying degrees, many retain fragmented class regions and mixed neighborhoods, particularly between affectively adjacent classes such as happy/excited and anger/frustration. SEDAL produces a more emotion-consistent manifold, with clearer local neighborhoods for sad, neutral, and excited utterances while preserving the proximity of related emotions. This structure is consistent with its design: MEMEA aligns unimodal and fused representations with momentum-evolving class anchors; STPL-PCR incorporates same- and cross-speaker transition patterns; AMPC stabilizes transition-related utterances under modality perturbation; and DSA-MTBD refines uncertain boundaries after the anchor geometry forms. SEDAL also achieves the highest 10-nearest-neighbor label purity (0.691) while maintaining high embedding trustworthiness (0.993). Together with the main results in Table 1 and the ablations in Table 2, these findings support the role of a structured multimodal emotion-anchor space in improving MERC.

Utter.	Speaker	Video	Dialogue	Label	CMERC	ECERC	SEDAL	Causal Analysis
u0	Rachel		No! I mean come on! This is a…	Anger	Anger	Anger	Anger	For u2: CMERC is pulled by Rachel's earlier anger and the interrogative form, treating the turn as another complaint. SEDAL instead models the adjacent cross-speaker transition from Joey's neutral uncertainty to Rachel's renewed positive curiosity, so the current anchor is attracted to Joy rather than Anger.
u1	Joey		I don't know.	Neutral	Neutral	Neutral	Neutral	
u2	Rachel		Is he romantic with her?	Joy	Anger	Joy	Joy	
u3	Joey		I don't know.	Neutral	Neutral	Neutral	Neutral	
u4	Rachel		Are they in love?	Joy	Joy	Joy	Joy	For u7: ECERC underestimates the sudden playful turn because the sentence is short and semantically underspecified after repeated neutral answers. SEDAL uses the cross-speaker Anger-to-Joy transition and multimodal anchor consistency to recognize Joey's playful reversal as Joy.
u5	Joey		I don't know.	Neutral	Neutral	Neutral	Neutral	
u6	Rachel		You don't know anything.	Anger	Anger	Anger	Anger	
u7	Joey		Ohh, I know one thing!	Joy	Joy	Neutral	Joy	

Figure 3: Case study on a MELD dialogue with repeated neutral replies and abrupt affective swings. Red predictions indicate baseline errors. By modeling cross-speaker transitions between Joey's neutral uncertainty and Rachel's changing affective state, SEDAL effectively captures the shift from anger to joy and achieves correct emotion prediction through speaker-aware transition prototypes.

4.5 Case Study

Figure 3 examines whether SEDAL captures cross-speaker emotional transitions by comparing it with CMERC and ECERC, which represent uncertainty calibration and evidence-cause reasoning, respectively. Red entries indicate utterances misclassified by at least one baseline, and the rightmost column summarizes each baseline failure and SEDAL's transition-based decision. In the dialogue, Rachel repeatedly questions Joey about Monica and Chandler, while Joey replies with the superficially neutral "I don't know." At u_2, CMERC predicts *Anger*, apparently propagating Rachel's earlier anger at u_0, whereas SEDAL captures her shift from Joey's neutral uncertainty to positive curiosity and predicts *Joy*. At u_7, ECERC labels Joey's underspecified "Ohh, I know one thing!" as *Neutral*; SEDAL recognizes his playful response to Rachel's preceding anger and keeps its anchor near the *Joy* prototype. This case illustrates the role of STPL-PCR. Rather than smoothing dialogue history or propagating dominant neighboring emotions, SEDAL represents each transition using adjacent multimodal anchors, their difference, and element-wise interaction, then indexes the resulting prototype by speaker relation. Transition-guided perturbation consistency further stabilizes transition-relevant predictions under weak or noisy unimodal evidence. These mechanisms help SEDAL resolve abrupt cross-speaker shifts. Appendix G.2 provides a same-speaker example.

Due to space limitations, more experimental analyses, including hyperparameter sensitivity analysis, error analysis, stability and robustness under modality damage, and additional case study details, are reported in Appendix G. Please refer to it for more experimental results and analyses that further validate the effectiveness of SEDAL.

5 Conclusion

In this paper, we propose SEDAL, a speaker-dependent emotional transition-aware framework for MERC. Its MEMEA module learns momentum-evolving class anchors that align unimodal and fused representations in a shared emotion-semantic space. STPL-PCR models same- and cross-speaker transitions with prototype learning and perturbation consistency, while DSA-MTBD distills reliable multi-teacher boundary knowledge into a single inference model. Experiments on IEMOCAP and MELD show that SEDAL achieves state-of-the-art performance, supporting speaker-aware transition anchor learning for MERC.

AI USE STATEMENT

In this work, generative artificial intelligence tools were used only to refine and polish the writing of this manuscript, including improving grammar, wording, clarity, and readability, and for suggesting alternative phrasings during revision. We have not used generative AI tools for generating experimental results, constructing or modifying datasets, designing the model architecture, deriving the mathematical formulations, or making autonomous scientific decisions. All technical ideas, methods, experiments, results, references, and conclusions were produced, checked, and approved by the authors.

ETHICS STATEMENT

This work builds upon two publicly released conversational emotion recognition benchmarks, IEMOCAP and MELD, and follows their standard, speaker-independent evaluation protocols. We introduce no new data collection, do not involve human subjects, and do not employ crowdsourcing. The datasets are used in the pre-processed, de-identified feature form released by their original authors, so no personally identifiable information is accessed or stored in our pipeline. All teacher models are trained on the training split only and selected exclusively by validation performance, with no access to test labels, which we believe mitigates concerns about data leakage and evaluation integrity. The teacher committee is used only during training and incurs no additional inference cost. While multimodal emotion recognition can in principle be applied to affect analysis of user-generated content, we discuss such applications only at the methodological level and do not deploy the proposed system on real-world user data; any downstream deployment should comply with applicable privacy regulations and consent norms. The authors declare no conflicts of interest or external sponsorship that could influence this work.

REPRODUCIBILITY STATEMENT

To facilitate the reproduction of SEDAL, we will release the source code, configuration files, and complete hyperparameter schedules. The model architecture, the training objective, and all module-specific losses are described in detail in the Method section and summarized in Algorithm 1. Implementation details—including the backbone encoders, optimization settings, learning rates, batch sizes, and training epochs on IEMOCAP and MELD—are provided in the Implementation Details subsection of Section 4. The teacher committee construction procedure, including the three diversity axes (training-trajectory, random-initialization, and objective-level inductive-bias variation), the marginal-diversity gate, and the validation-based member weighting, is specified in the DSA-MTBD subsection. Both benchmarks used in this paper are publicly available, and our data processing follows their standard released pipelines. All results reported in the paper are averaged over five runs with standard deviations.

REFERENCES

Wei Ai, Yuntao Shou, Tao Meng, and Keqin Li. DER-GCN: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):4908–4921, 2025. doi: 10.1109/TNNLS.2024.3367940.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.

Xinyi Che, Wenbo Wang, Yuanbo Hou, Mingjie Xie, Qijun Zhao, and Jian Guan. Angle-optimized partial disentanglement for multimodal emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 17(3):2612–2624, 2026. doi: 10.1109/TAFFC.2026.3683470.

Xuping Chen and Wuzhen Shi. Dynamic interactive bimodal hypergraph networks for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1256–1264, 2025. doi: 10.1609/aaai.v39i2.32114.

Yijing Dai, Yingjian Li, Dongpeng Chen, Jinxing Li, and Guangming Lu. Multimodal decoupled distillation graph neural network for emotion recognition in conversation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9910–9924, 2024. doi: 10.1109/TCSVT.2024.3405406.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 154–164, 2019.

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*, pages 1122–1131. ACM, 2020. doi: 10.1145/3394171.3413678.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052. Association for Computational Linguistics, 2021a. doi: 10.18653/v1/2021.acl-long.547.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9747397.

Guanyu Hu, Dimitrios Kollias, and Xinyu Yang. Grounding emotion recognition with visual prototypes: Vega - revisiting clip in merc. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5667–5676. ACM, 2025. doi: 10.1145/3746027.3755340.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 5666–5675, 2021b.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.306.

Jae-Jun Lee and Sung Whan Yoon. Can one modality model synergize training of other modality models? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b7acc4653f48f54bba55f55d2cd10417-Abstract-Conference.html.

Fenglin Li, Mengyin Wang, Yanping Yang, Fan Zhang, Zhaofeng Chen, and Houjie Li. BiSSM-CL: Enhanced multimodal emotion recognition in conversation with bidirectional state space model and collaborative learning. *Information Fusion*, 135:104455, 2026. doi: 10.1016/j.inffus.2026.104455.

Jiang Li, Xiaoping Wang, and Zhigang Zeng. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8786–8803, 2025a. doi: 10.1109/TPAMI.2025.3581236.

Jie Li, Shifei Ding, Lili Guo, and Xuan Li. Multi-modal anchor gated transformer with knowledge distillation for emotion recognition in conversation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 8141–8149, 2025b. doi: 10.24963/ijcai.2025/905.

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*, pages 6631–6640, 2023.

Zheng Lian, Bin Liu, and Jianhua Tao. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000, 2021. doi: 10.1109/TASLP.2021.3049898.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26:776–788, 2024. doi: 10.1109/TMM.2023.3271019.

Anh-Tuan Mai, Cam-Van Thi Nguyen, and Duc-Trong Le. Divide and refine: Enhancing multimodal representation and explainability for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2700–2709, 2026. doi: 10.1109/WACV61042.2026.00265.

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

Tao Meng, Weilun Tang, Yuntao Shou, Yilong Tan, Jun Zhou, Wei Ai, and Keqin Li. Dynamic fusion-aware graph convolutional neural network for multimodal emotion recognition in conversations. *Knowledge-Based Systems*, 343:115846, 2026. doi: 10.1016/j.knosys.2026.115846.

Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.937.

Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. Ada2I: Enhancing modality balance for multimodal conversational emotion recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9330–9339. ACM, 2024. doi: 10.1145/3664647.3681648.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1050.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.123.

Jiandong Shi, Ming Li, Lu Bai, Feilong Cao, Ke Lu, and Jiye Liang. MATCH: Modality-calibrated hypergraph fusion network for conversational emotion recognition. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6164–6172, 2025. doi: 10.24963/ijcai.2025/686.

Yuntao Shou, Huan Liu, Xiangyong Cao, Deyu Meng, and Bo Dong. A low-rank matching attention based cross-modal feature fusion method for conversational emotion recognition. *IEEE Transactions on Affective Computing*, 16(4):1177–1189, 2025.

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*, pages 6558–6569. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1656.

Geng Tu, Tian Xie, Bin Liang, Hongpeng Wang, and Ruifeng Xu. Adaptive graph learning for multimodal conversational emotion detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19089–19097, 2024a. doi: 10.1609/aaai.v38i17.29876.

Geng Tu, Feng Xiong, Bin Liang, Hui Wang, Xi Zeng, and Ruifeng Xu. Multimodal emotion recognition calibration in conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9621–9630. ACM, 2024b. doi: 10.1145/3664647.3681515.

Yusong Wang, Xuanye Fang, Huifeng Yin, Dongyuan Li, Guoqi Li, Qi Xu, Yi Xu, Shuai Zhong, and Mingkun Xu. BIG-FUSION: Brain-inspired global-local context fusion framework for multimodal emotion recognition in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1574–1582, 2025. doi: 10.1609/aaai.v39i2.32149.

Zijian Yi, Ziming Zhao, Zhishu Shen, and Tiehua Zhang. Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4341–4348. ACM, 2024. doi: 10.1145/3664647.3681633.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797, 2021. doi: 10.1609/aaai.v35i12.17289.

Tao Zhang and Zhenhua Tan. 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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.102.

Xiaofei Zhu, Yang Jiang, Xiaoyang Liu, and Yihao Zhang. Cross-modality multiband differential conditional diffusion for multimodal emotion recognition in conversation. *Knowledge-Based Systems*, 340:115594, 2026. doi: 10.1016/j.knosys.2026.115594.

APPENDIX

In this appendix, we provide the supplementary materials that support the main text. Appendix A summarizes the notation used throughout the paper, including the updated symbols for multimodal branches, emotion anchors, speaker-aware transitions, perturbation consistency, and multi-teacher distillation. Appendix B gives the training pseudocode of SEDAL. Appendices C and D describe the benchmark datasets and evaluation metrics, Appendix E summarizes the compared baselines, Appendix F reports implementation details, and Appendix G provides additional experimental analyses, including robustness under modality damage, extra case studies, hyperparameter sensitivity, and error analysis. The detailed contents are organized as follows:

A Default Notation 14

B Algorithm Pseudocode 19

C Datasets Description 19

C.1 IEMOCAP 19

C.2 MELD 20

C.3 Selection Rationale 20

D Evaluation Metrics 20

D.1 Accuracy (ACC) 20

D.2 Weighted F1-score (W-F1) 20

E Baselines 21

F Implementation Details 22

G Additional Experimental Analysis 22

G.1 Stability and Robustness Analysis 22

G.2 More Case Study 23

G.3 Hyperparameter Sensitivity Analysis 24

G.4 Error Analysis 25

A Default Notation

In this appendix we summarize the notation used throughout the paper. Bold lowercase letters denote vectors, bold uppercase letters denote matrices, and calligraphic letters denote sets, prototype banks, and distributions. Unless otherwise stated, all vectors are column vectors, $\|\cdot\|$ is the Euclidean norm, and all logarithms are natural logarithms. Symbols that appear only once are defined at their first occurrence in the main text.

Dialogue and Data

$\mathcal{D}$	A dialogue, $\mathcal{D} = \{(u_t, s_t, y_t)\}_{t=1}^{T}$
u_t	The t-th utterance in a dialogue
s_t	The speaker identity of the t-th utterance
y_t	The emotion label of the t-th utterance
T	The number of utterances in a dialogue
C	The number of emotion classes
$\mathcal{C}$	The set of emotion classes, $\mathcal{C} = \{1, \ldots, C\}$

$\mathcal{B}$	A mini-batch of dialogues and their valid utterances
$\mathcal{M}$	The set of unimodal streams, $\mathcal{M} = \{T, A, V\}$ for text, audio, and vision
$\mathcal{M}^+$	The set of prediction branches, $\mathcal{M}^+ = \{T, A, V, M\}$, where M denotes the fused multimodal branch
$\delta_{i,t}$	A validity indicator; $\delta_{i,t} = 1$ if u_t in dialogue i is a valid non-padding utterance and 0 otherwise
$\mathbf{x}_t^m$	The extracted input descriptor of modality m for u_t

Spaces, Encoders, and Representations

$\mathbb{R}$	The set of real numbers
$\mathbb{R}^d$	The d-dimensional real vector space
d_h	The hidden dimension of the backbone representations
d_z	The dimension of the multimodal emotion-anchor space
d_s	The dimension of the speaker embedding
d_{tr}	The dimension of the transition-anchor space
Δ^{C-1}	The probability simplex over C emotion classes
$F_{enc}^m(\cdot)$	The modality-specific input encoder of modality m
$\mathrm{Conv1D}_m(\cdot)$	The 1-D convolutional projection for modality m
$\mathbf{e}_t^m$	The utterance-level feature of modality m, $\mathbf{e}_t^m = \mathrm{Conv1D}_m(F_{enc}^m(\mathbf{x}_t^m))$
$\mathbf{E}^m$	The projected descriptor sequence of modality m over a dialogue
$\mathbf{H}^m$	The context-aware representation sequence of modality m
$\mathbf{h}_t^m$	The context-aware representation of u_t under modality m
$\boldsymbol{\pi}_t$	The modality gating vector of the backbone fusion for u_t
$\mathbf{h}_t^M$	The gated fused multimodal representation of u_t
$W_f, \mathbf{b}_f$	The weight matrix and bias of the gated fusion layer
$W_o^m, \mathbf{b}_o^m$	The weight matrix and bias of the branch-wise classifier of branch m
$\mathbf{o}_t^m$	The branch-wise logit vector of branch m for u_t
$\mathbf{p}_t^m$	The branch-wise emotion probability, $\mathbf{p}_t^m = \mathrm{softmax}(\mathbf{o}_t^m)$
$\hat{y}_t$	The final emotion prediction of u_t, produced from the fused branch $\mathbf{p}_t^M$
$\mathbf{e}_{s_t}$	The speaker embedding of s_t

Anchors and Prototypes

$f_m(\cdot)$	The projection head that maps branch m into the anchor space
$g(\cdot)$	The gating network in the anchor space
$\mathbf{z}_t^m$	The normalized anchor-space representation of branch m for u_t
$\boldsymbol{\alpha}_t$	The modality weight vector of the multimodal anchor for u_t
$\mathbf{r}_t$	The multimodal anchor feature of u_t
$\mathbf{P}$	The class anchor bank, $\mathbf{P} \in \mathbb{R}^{C \times d_z}$
$\mathbf{p}_c$	The emotion prototype of class c in the anchor bank
$\bar{\mathbf{r}}_c$	The batch-wise anchor center of class c
μ_{ma}	The momentum coefficient of the MEMEA updates
τ_{ma}	The temperature of the anchor prototype logits
$\ell_{t,c}^m$	The anchor-prototype logit of branch m for class c at u_t
$\boldsymbol{\ell}_t^m$	The anchor-prototype logit vector of branch m for u_t
$\boldsymbol{\ell}_t^{anc}$	The clean anchor-prototype logit vector used as the AMPC prototype target, instantiated by the fused anchor branch

Transitions and Perturbations

$\mathbf{r}_t^s$	The speaker-conditioned anchor feature, $\mathbf{r}_t^s = [\mathbf{r}_t; \mathbf{e}_{s_t}]$
$\Delta\mathbf{r}_t^s$	The speaker-conditioned anchor difference, $\Delta\mathbf{r}_t^s = \mathbf{r}_t^s - \mathbf{r}_{t-1}^s$
$\mathbf{q}_t$	The transition input of the adjacent pair (u_{t-1}, u_t)
$f_{tr}(\cdot)$	The transition projection head
$\mathbf{d}_t$	The normalized transition representation of u_t
$\mathcal{T}$	The set of valid adjacent utterance pairs in a mini-batch
ρ_t	The speaker-relation indicator; $\rho_t = 0$ for same-speaker and 1 for cross-speaker transitions
$\mathbf{A}$	The speaker-aware transition prototype bank, $\mathbf{A} \in \mathbb{R}^{2 \times C \times d_{tr}}$
$\mathbf{a}_{\rho,c}$	The transition prototype of speaker relation ρ and class c
$\bar{\mathbf{d}}_{\rho,c}$	The batch-wise transition center of relation ρ and class c
μ_{tr}	The momentum coefficient of the transition-prototype updates
τ_{tr}	The temperature of the transition prototype logits
$\eta_{t,c}$	The transition-prototype logit for class c at u_t, computed under the observed speaker relation
κ_t	The modality selected for perturbation at u_t

$\tilde{\mathbf{h}}_t^m$	The perturbed representation of branch m
$\tilde{\mathbf{z}}_t^m$	The perturbed anchor-space projection of branch m
$\tilde{\mathbf{r}}_t$	The perturbed multimodal anchor feature of u_t
$\tilde{\boldsymbol{\ell}}_t^m$	The perturbed anchor-prototype logit vector of branch m
$\tilde{\boldsymbol{\ell}}_t^{anc}$	The perturbed anchor-prototype logit vector used in AMPC
$\tilde{\mathbf{o}}_t^M$	The perturbed fused-branch logits of u_t
$\mathcal{P}$	The set of actually perturbed utterances in a mini-batch

Multi-Teacher Distillation

$\mathcal{K}$	The teacher committee
$\mathcal{K}_{pool}$	The candidate pool of training-only teacher models
$\mathcal{V}$	The validation split
$\mathbf{p}_{k,t}^m$	The probability vector predicted by teacher member k and branch m for utterance u_t
$\mathbf{p}_{k,m}(u_t)$	Equivalent functional notation for the teacher distribution of member k and branch m
$\mathbf{p}_{new,\mathrm{M}}(u_t), \mathbf{p}_{j,\mathrm{M}}(u_t)$	The fused-branch validation predictions used to compute teacher diversity
α_k	The aggregation weight of teacher member k
β_m	The aggregation weight of teacher branch m
$\mathrm{Acc}_k(\mathcal{V})$	The validation accuracy of teacher member k
s_k	The selection score of teacher member k
d_{new}	The marginal diversity of a new teacher candidate
d_{min}	The diversity acceptance threshold
τ_{val}	The validation-based teacher selection threshold
η	The performance-proportional weighting power of the teacher aggregation
$\mathbf{Q}_t$	The committee (teacher) distribution for u_t
$Q_{t,c}$	The teacher probability of class c for u_t
v_t	The reliability mask of the teacher target for u_t
τ_{conf}	The teacher confidence threshold
γ	The confidence-weighting power
r_t	The student-adaptive focus weight for u_t
$p_s^M(\hat{c}_t \mid u_t)$	The fused-branch student probability assigned to the teacher top class $\hat{c}_t$
ρ	The student-focus power
$\hat{c}_t$	The teacher top-1 class of u_t
e_0	The epoch at which distillation is activated
w	The distillation warmup length in epochs

λ_{kd}^{max}	The maximum distillation weight
ε	The label-smoothing coefficient of the teacher distribution
$\bar{\mathbf{Q}}_t$	The smoothed teacher distribution for u_t
T_{kd}	The distillation temperature

Losses and Objectives

$\mathrm{CE}(\cdot,\cdot)$	The cross-entropy loss
$\mathrm{KL}(\cdot\|\cdot)$	The Kullback–Leibler divergence
$\mathcal{U}$	The set of valid utterances in a mini-batch
$\mathcal{L}_{cls}$	The backbone classification loss
ω_m	The branch weight of the MEMEA loss
$\mathcal{L}_{ma}$	The MEMEA anchor-prototype loss
$\mathcal{L}_{tr}$	The transition-prototype loss
τ_{pc}	The temperature of the perturbation consistency losses
$\mathcal{L}_{pc}^{anc}$	The anchor-feature consistency loss
$\mathcal{L}_{pc}^{proto}$	The anchor-prototype-logit consistency loss
$\mathcal{L}_{pc}^{cls}$	The classifier-logit consistency loss
$\beta_{anc}, \beta_{proto}, \beta_{cls}$	The weights of the three perturbation consistency terms
$\mathcal{L}_{pc}$	The perturbation consistency loss
$\mathcal{L}_{kd}$	The multi-teacher boundary distillation loss
$\lambda_{ma}, \lambda_{tr}, \lambda_{pc}, \lambda_{kd}(e)$	The loss weights for MEMEA, STPL-PCR, AMPC, and deferred distillation
$\mathcal{L}$	The overall training objective
$\theta, \theta^{\star}$	The model parameters and the learned optimal parameters
e	The training epoch
E	The total number of training epochs

Operators and Indexing

$[\cdot;\cdot]$	Vector concatenation
$\odot$	Element-wise (Hadamard) product
$\mathbb{I}[\cdot]$	The indicator function; 1 if the condition is true and 0 otherwise
$\mathrm{softmax}(\cdot)$	The softmax function
$\|\cdot\|$	The ℓ_2 norm
$\mathrm{sg}(\cdot)$	The stop-gradient operator
$\mathrm{JS}(\cdot,\cdot)$	The Jensen–Shannon divergence

$\mathrm{Bernoulli}(p)$	The Bernoulli distribution with success probability p
$\mathrm{Uniform}(\mathcal{M})$	The uniform distribution over the modalities $\mathcal{M}$
ϵ	A small constant used to avoid division by zero
i	The dialogue index within a mini-batch
t	The utterance index within a dialogue
m	A modality or branch index; $m \in \mathcal{M}$ for unimodal streams and $m \in \mathcal{M}^+$ when the fused branch is included
c	An emotion-class index, $c \in \mathcal{C}$
k	A teacher-member index, $k \in \mathcal{K}$

B Algorithm Pseudocode

Algorithm 1 summarizes the SEDAL training procedure. The ordering is deliberate: classification and MEMEA first establish the anchor space, STPL-PCR then learns speaker-aware transition prototypes and transition-guided robustness within that space, and DSA-MTBD finally compresses reliable teacher boundary information after the student has become stable enough to benefit from it.

Algorithm 1: The Proposed SEDAL Framework.

Input : Training mini-batches $\mathcal{B}$ with labels and validity masks $\{\delta_{i,t}\}$; SEDAL hyperparameters; optional teacher committee distributions $\{\mathbf{Q}_{i,t}\}$

Output : Emotion probabilities $\{\mathbf{p}_{t,c}^{all}\}$, predictions $\{\hat{y}_t\}$, and trained parameters $\theta^\star$

Initialization: class anchor bank $\mathbf{P}$ and transition bank $\mathbf{A}$;
for *each training epoch* $e = 1, \ldots, E$ **do**
  for *each mini-batch* $\mathcal{B}$ **do**
    Encode multimodal dialogues into branch features $\{\mathbf{h}_{i,t}^m\}_{m\in\mathcal{M}^+}$ and logits $\{\mathbf{o}_{i,t}^m\}_{m\in\mathcal{M}^+}$; compute $\mathcal{L}_{cls}$;
    MEMEA: construct anchors $\{\mathbf{r}_{i,t}\}$, update $\mathbf{P}$, and compute $\mathcal{L}_{ma}$;
    STPL: form $\mathcal{T}$, speaker-conditioned features $\{\mathbf{d}_{i,t}\}$, and relations $\{\rho_{i,t}\}$;
    if $\mathcal{T} \neq \emptyset$ **then**
      Update transition bank $\mathbf{A}$ and compute $\mathcal{L}_{tr}$;
    end
    AMPC: build transition mask, sample modality perturbations, recompute perturbed outputs, and compute $\mathcal{L}_{pc}$;
    if $\lambda_{kd}(e) > 0$ **then**
      DSA-MTBD: form $v_{i,t}$, $r_{i,t}$, and smoothed teachers from $\{\mathbf{Q}_{i,t}\}$; compute $\mathcal{L}_{kd}$;
    end
    Update θ by minimizing $\mathcal{L}$;
  end
end
Infer with the single student model to obtain $\{\mathbf{p}_{t,c}^{all}\}$ and $\{\hat{y}_t\}$;
return $\theta^\star$, $\{\mathbf{p}_{t,c}^{all}\}$, and $\{\hat{y}_t\}$;

C Datasets Description

C.1 IEMOCAP

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) (Busso et al., 2008) is a widely used benchmark for conversational emotion recognition. It contains approximately 12 hours of dyadic conversations performed by ten professional actors (five female and five male), recorded across five

sessions with two actors per session. Each utterance is annotated with one of six emotion categories, namely angry, excited, frustrated, happy, neutral, and sad, together with dimensional annotations including valence, arousal, and dominance. For every utterance, IEMOCAP provides synchronized multimodal streams, covering audio, video, manually transcribed text, and motion-capture recordings. In our experiments, we follow the standard speaker-independent evaluation protocol of MERC, in which two speakers (one female and one male) are held out for testing so that no speaker appears in both the training and test sets, which measures the generalization of the model to unseen speakers.

C.2 MELD

MELD (Multimodal EmotionLines Dataset) (Poria et al., 2019) is a multi-party conversational emotion recognition benchmark derived from the TV series *Friends*. It contains 1,433 dialogues and 13,708 utterances extracted from 13 episodes, and each utterance is annotated with one of seven emotion categories, namely anger, disgust, fear, joy, neutral, sadness, and surprise, as well as a sentiment label (positive, negative, or neutral). MELD provides aligned textual, acoustic, and visual modalities for each utterance, where the visual stream is extracted from the video frames of the corresponding speaker. The dataset is officially split into 1,039 dialogues for training, 114 for validation, and 280 for testing. Compared with IEMOCAP, MELD involves multi-party interactions, a larger number of speakers, and a larger emotion label space, which makes it complementary for evaluating the generalization of the proposed SEDAL framework.

C.3 Selection Rationale

We select IEMOCAP and MELD for the following reasons. First, both benchmarks provide synchronized textual, acoustic, and visual signals for every utterance, which is a prerequisite for multimodal emotion recognition in conversations. Second, they cover complementary conversational settings: IEMOCAP consists of dyadic conversations with six emotion categories and a limited number of speakers, whereas MELD contains multi-party dialogues with seven emotion categories, enabling us to evaluate the proposed speaker-aware transition prototype learning and transition-guided robustness in both two-party and multi-party scenarios. Third, both datasets are publicly available and widely adopted in recent MERC studies, which allows direct and fair comparison with state-of-the-art baselines under the same protocol.

D Evaluation Metrics

D.1 Accuracy (ACC)

Accuracy is the proportion of correctly predicted utterances among all valid utterances. Given a test set of N utterances with ground-truth labels y_i and model predictions $\hat{y}_i$, the accuracy is computed as

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i], \tag{18}$$

where $\mathbb{I}[\cdot]$ is the indicator function, which equals 1 if the prediction matches the ground truth and 0 otherwise.

D.2 Weighted F1-score (W-F1)

To account for the class imbalance that commonly exists in conversational emotion benchmarks, we additionally report the weighted F1-score (W-F1), which computes the F1-score of each emotion class and averages these scores weighted by the number of ground-truth samples of each class. For a class c, let TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives, respectively. The precision and recall of class c are defined as

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \tag{19}$$

and the F1-score of class c is the harmonic mean of precision and recall:

$$F1_c = \frac{2P_cR_c}{P_c + R_c}. \tag{20}$$

The weighted F1-score is then obtained by averaging the per-class F1-scores with weights proportional to the class support:

$$\text{W-F1} = \sum_{c \in \mathcal{C}} \frac{N_c}{N} F1_c, \tag{21}$$

where N_c is the number of utterances whose ground-truth label is class c, $N = \sum_{c \in \mathcal{C}} N_c$ is the total number of utterances, and $\mathcal{C}$ is the set of emotion classes. Compared with the macro-averaged F1-score, which treats all classes equally, W-F1 assigns larger weight to frequent classes and therefore better reflects the overall recognition quality on imbalanced conversational emotion benchmarks.

E BASELINES

We compare SEDAL with 18 recent and representative MERC baselines. The compared methods are summarized as follows.

- **SDT** (Ma et al., 2024): It uses Transformer blocks to capture intra-modal and cross-modal interactions, and introduces self-distillation so that unimodal branches can learn from stronger multimodal representations.
- **D^{2}GCN** (Dai et al., 2024): It builds a graph neural network with decoupled distillation, separating emotion-relevant knowledge from redundant multimodal information to improve conversational reasoning.
- **HAUCL** (Yi et al., 2024): It formulates multimodal dialogue fusion as hypergraph learning, where a hypergraph autoencoder models high-order utterance relations and contrastive learning regularizes uncertain multimodal representations.
- **AdaIGN** (Tu et al., 2024a): It learns adaptive interactive graphs for conversational emotion detection by selecting informative nodes and edges, allowing the model to focus on useful contextual dependencies instead of using a fixed graph.
- **CMERC** (Tu et al., 2024b): It treats MERC as a calibration problem and combines curriculum-style training, supervised contrastive learning, and confidence constraints to reduce overconfident predictions on difficult samples.
- **DER-GCN** (Ai et al., 2025): It constructs dialogue- and event-aware graphs to model speaker relations, temporal dependencies, and event-level emotion cues in multimodal conversations.
- **MAGTKD** (Li et al., 2025b): It combines an anchor-gated Transformer with knowledge distillation, using stronger multimodal or textual signals to guide weaker modality branches.
- **ECERC** (Zhang and Tan, 2025): It introduces an evidence-cause attention network that jointly models emotion evidence and contextual causes, so that emotion prediction can attend to both local cues and causal dialogue context.
- **MATCH** (Shi et al., 2025): It uses modality-calibrated hypergraph fusion, calibrating entity-level modality representations and then modeling high-order emotion relations with hypergraph structures.
- **GraphSmile** (Li et al., 2025a): It jointly learns graph structures and sentiment dynamics, capturing both multimodal dependencies and the evolving affective tendency in a dialogue.
- **BIG-FUSION** (Wang et al., 2025): It adopts a brain-inspired global-local fusion framework, combining global dialogue context with local contextual windows to model emotion dynamics at different ranges.
- **DIB-HGCN** (Chen and Shi, 2025): It builds dynamic interactive bimodal hypergraphs, separating monologic and dialogic relations to capture speaker-specific and cross-speaker dependencies.
- **VEGA** (Hu et al., 2025): It revisits CLIP for MERC by grounding visual emotion recognition with visual prototypes, aligning utterance representations with emotion-level visual anchors.
- **DF-GCN** (Meng et al., 2026): It designs a dynamic fusion-aware graph convolutional network, using category-aware graph reasoning to adapt multimodal fusion to different emotional states.

- **CMDCD** (Zhu et al., 2026): It applies cross-modality multiband differential conditional diffusion, denoising and enhancing multimodal features across frequency bands before emotion classification.
- **AO-FL** (Che et al., 2026): It performs angle-optimized partial disentanglement, separating shared and modality-specific factors while constraining their geometry for more reliable fusion.
- **DnR** (Mai et al., 2026): It follows a divide-and-refine strategy that decomposes multimodal representations into different information components and refines them with component-specific objectives.
- **BiSSM-CL** (Li et al., 2026): It combines a bidirectional state-space model with collaborative learning, using sequence modeling and mutual guidance among modality branches to strengthen conversational context modeling.

F Implementation Details

We implemented the model in PyTorch. Each utterance is encoded with textual, acoustic, and visual descriptors, then projected into a shared Transformer space with speaker embeddings and modality-specific interaction branches. CLIP ViT-B/32 is used only to construct fixed visual emotion anchors, and its encoder is kept frozen. We optimize SEDAL with AdamW using a learning rate close to 1×10^{-4}. The batch sizes are 9 and 32 on IEMOCAP and MELD, and the training epochs are 400 and 52, respectively. The hidden dimensions and attention heads are set to 768/4 on IEMOCAP and 1280/8 on MELD. The peak weights of the auxiliary objectives are set to $\lambda_{ma}^{\max} = 0.12$ with a 5-epoch warm-up, $\lambda_{tr}^{\max} = 0.005$ with an 8-epoch warm-up, and $\lambda_{pc}^{\max} = 0.03$ with a 5-epoch warm-up, so that the multimodal anchors stabilize before the speaker-aware transition prototypes and perturbation consistency are introduced. The deferred distillation weight is set to $\lambda_{kd}^{\max} = 0.25$, and it is activated from epoch $e_0 = 35$ and linearly warmed up over $w = 8$ epochs. Within the AMPC objective, the anchor-feature, prototype-logit, and classifier-logit consistency terms are weighted by $\beta_{anc} = 1.0$, $\beta_{proto} = 1.0$, and $\beta_{cls} = 0.5$, respectively, and the shared perturbation-consistency temperature is set to $\tau_{pc} = 2.0$. For the teacher committee, we train candidates along the training-trajectory, random-initialization, and objective-level inductive-bias axes; the marginal-diversity threshold is set to $d_{min} = 0.5\,\bar{\mathrm{JS}}$, where $\bar{\mathrm{JS}}$ is the mean pairwise Jensen–Shannon divergence among already-admitted pool members. All experiments were conducted on an NVIDIA RTX 4090 GPU.

G Additional Experimental Analysis

G.1 Stability and Robustness Analysis

We evaluate whether the perturbation-consistency part of STPL-PCR improves stability under modality damage. Table 10 and Figure 4 show that transition-guided AMPC consistently improves over the variant without AMPC across unimodal, bimodal, and dropout settings on both IEMOCAP and MELD. On IEMOCAP, the W-F1 gain ranges from +0.59 to +2.32 points and the Acc gain from +0.75 to +2.19 points. On MELD, the corresponding gains are +0.86 to +1.85 W-F1 and +0.87

Table 10: Stability and robustness results (%) with and without transition-guided AMPC on IEMOCAP and MELD.

Modality	w/ AMPC (Transition-guided)				w/o AMPC			
	IEMOCAP		MELD		IEMOCAP		MELD	
	Acc	W-F1	Acc	W-F1	Acc	W-F1	Acc	W-F1
Text	68.87	69.23	66.74	65.79	68.12	68.64	65.87	64.93
Audio	63.45	64.07	50.56	43.25	62.69	63.3	48.82	41.59
Visual	44.93	45.48	49.43	36.02	43.74	44.31	48.13	34.66
Text + Audio	73.12	73.24	67.31	66.94	71.83	72.09	66.15	65.64
Text + Visual	70.51	70.76	67.36	66.72	69.34	69.55	66.28	65.57
Audio + Visual	64.27	64.59	50.24	42.36	62.46	63.08	49.02	41.2
dropout T+A+V 0.5	70.56	70.84	61.12	60.76	68.37	68.52	59.53	58.91

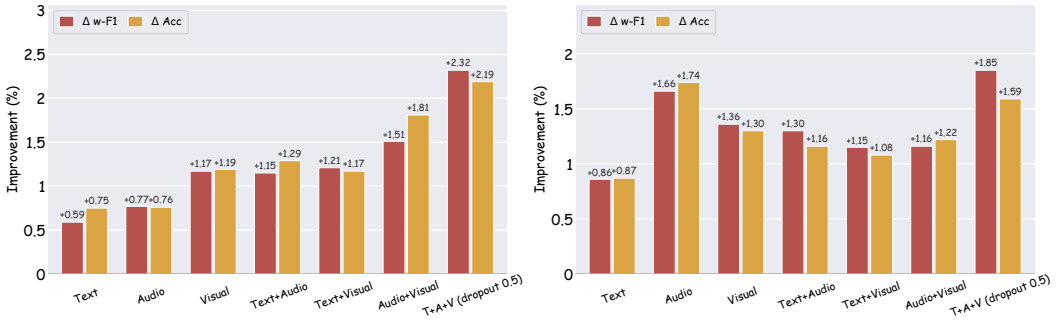


Figure 4: Performance gain of transition-guided AMPC over the variant without AMPC under different modality availability and dropout settings. The left panel reports IEMOCAP results, and the right panel reports MELD results.

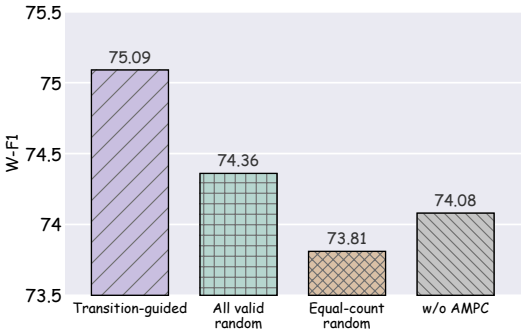


Figure 5: IEMOCAP W-F1 comparison of different AMPC perturbation-position strategies. Transition-guided perturbation selects utterances with valid transition context, all-valid random samples from all valid utterances, and equal-count random uses the same perturbation count as the transition-guided setting.

to +1.74 Acc. The largest gains appear under explicit tri-modal dropout, where AMPC improves IEMOCAP by +2.32 W-F1 and +2.19 Acc and MELD by +1.85 W-F1 and +1.59 Acc. This pattern is consistent with the objective of AMPC: damaged views are encouraged to preserve the clean anchor feature, anchor-prototype distribution, and classifier prediction at transition-related positions.

Figure 5 further isolates where AMPC should apply perturbations on IEMOCAP. The proposed transition-guided strategy perturbs only current utterances with valid preceding context, matching the positions used by speaker-aware transition prototype learning. It reaches 75.09 W-F1, higher than perturbing all valid utterances at random (74.36), perturbing an equal number of randomly selected utterances (73.81), and removing AMPC (74.08). This comparison separates the effect of perturbation location from the amount of perturbation, showing that AMPC works best when it regularizes transition-relevant utterances rather than spreading modality damage uniformly over the dialogue.

G.2 More Details for Case Study

To examine whether SEDAL captures same-speaker emotional transitions, we present the qualitative case study in Figure 6. After expressing sadness over a rejection, Chandler gradually shifts the conversation toward requesting help for his actor friend. At u_5, CMERC predicts *Surprise* for the hesitant utterance "Well, I umm, I mean this is just off the top of my head...", likely because its discourse structure resembles a reveal. SEDAL instead recognizes it as a neutral preface by relating the current turn to Dana's preceding supportive joy and the local transition context. At u_6, CMERC predicts *Joy*, possibly emphasizing the positive goal of securing an audition, while ECERC predicts

Utter.	Speaker	Video	Dialogue	Label	CMERC	ECERC	SEDAL	Causal Analysis
u0	Dana		Of course!	Joy	Joy	Joy	Joy	For u5: CMERC over-reacts to the hesitation and the reveal-like discourse structure, so it predicts Surprise. SEDAL conditions the turn on Dana's preceding supportive Joy and Chandler's request setup, where the current anchor remains a neutral preface before the real emotional escalation.
u1	Chandler		Well, it just hurts so bad. I uh…	Sadness	Sadness	Sadness	Sadness	
u2	Dana		Ohh, I'm sorry.	Sadness	Sadness	Sadness	Sadness	
u3	Chandler		Well maybe there is one thing you can do.	Neutral	Neutral	Neutral	Neutral	
u4	Dana		What?! Anything! Anything!	Joy	Joy	Joy	Joy	For u6: Both baselines are distracted by the polite request and positive goal of getting an audition. SEDAL separates this same-speaker Neutral-to-Fear transition from the previous preface and captures Chandler's anxious cue ('would kill me') through the transition-anchor difference.
u5	Chandler		Well, I umm, I mean this is just off the top of my head now, umm but I have this friend.	Neutral	Surprise	Neutral	Neutral	
u6	Chandler		This actor friend and he would kill me if he thought I was doing this umm, but umm would it be possible for him to get an audition for your movie say on Thursday?	Fear	Joy	Neutral	Fear	
u7	Dana		Absolutely! But you-you would really feel better about me rejecting you if your actor friend can audition for my movie?	Surprise	Surprise	Surprise	Surprise	

Figure 6: Case study on a request-driven MELD dialogue after rejection. Red predictions indicate baseline errors. The proposed SEDAL distinguishes a neutral request preface from a subsequent same-speaker transition into fear by comparing adjacent multimodal emotion anchors and modeling the transition direction with speaker-aware prototypes, leading to more reliable emotion prediction.

Neutral based on the polite request form. SEDAL correctly identifies *Fear* by capturing Chandler's shift from a neutral preface to an anxious request, reinforced by the cue "he would kill me."

This case illustrates the role of STPL-PCR in modeling emotional continuity within the same speaker. SEDAL forms transition features from adjacent multimodal anchors, their difference, and element-wise interaction, then indexes the resulting prototypes by same-speaker relations. This representation models how the current emotion emerges from the preceding conversational state, while transition-guided perturbation consistency reduces sensitivity to weak or noisy unimodal evidence. Together, these mechanisms help SEDAL resolve the highlighted errors during same-speaker emotional escalation.

G.3 Hyperparameter Sensitivity Analysis

We further examine the sensitivity of SEDAL to the main training hyperparameters on IEMOCAP. Figure 7 reports the weighted F1 and weighted accuracy under different settings for the MEMEA anchor objective, the STPL transition objective, and the AMPC perturbation consistency objective. Overall, SEDAL is not overly sensitive within a reasonable range, but each module benefits from a balanced schedule. MEMEA performs best at $\lambda_{anchor} = 0.12$ with a 5-epoch warm-up, showing that the anchor loss should be strong enough to align class-level semantics but not so large that it dominates task supervision. STPL reaches its best result with $\lambda_{trans} = 0.005$ and an 8-epoch warm-up, which supports introducing transition prototypes after the multimodal anchor geometry has started to stabilize. For AMPC, $\lambda_{AMPC} = 0.03$ and $p_{perturb} = 0.35$ provide the best trade-off: weaker perturbation underuses the robustness signal, while stronger perturbation can damage emotion-discriminative cues. These trends are consistent with the staged design of SEDAL, where anchor learning first establishes the semantic space, transition learning then shapes speaker-aware dynamics, and perturbation consistency regularizes difficult transition-related utterances without dominating supervised learning.

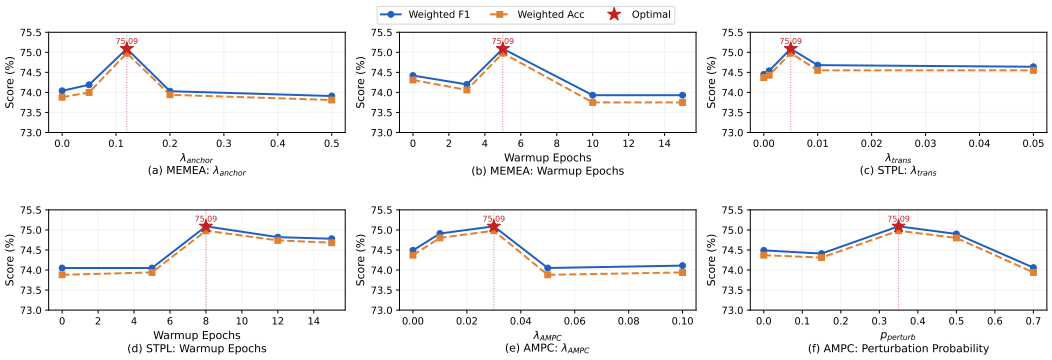
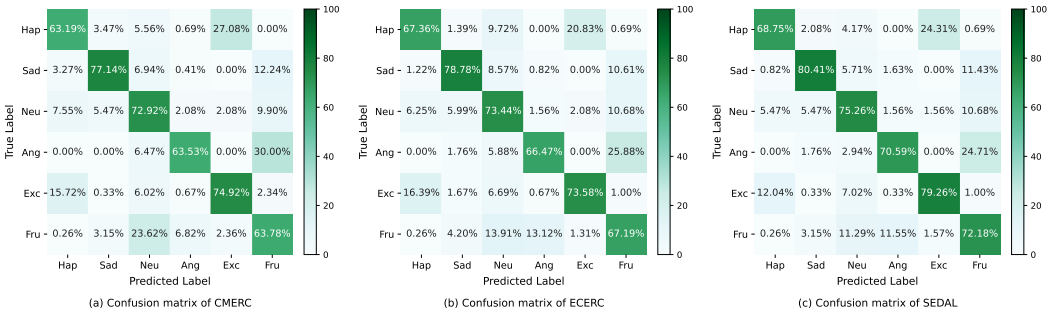
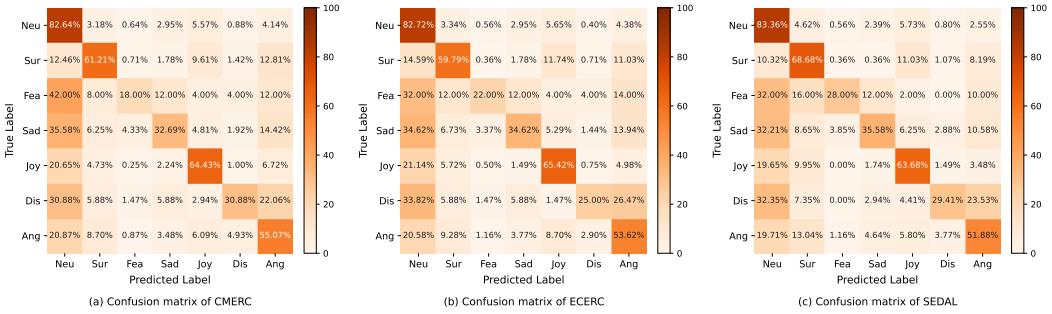


Figure 7: Hyperparameter sensitivity analysis experiments conducted on IEMOCAP.



(a) CMERC (b) ECERC (c) SEDAL

Figure 8: Comparison of confusion matrices among CMERC, ECERC, and SEDAL for the IEMOCAP dataset.



(a) CMERC (b) ECERC (c) SEDAL

Figure 9: Comparison of confusion matrices among CMERC, ECERC, and SEDAL for the MELD dataset.

G.4 Error Analysis

Figure 8 and Figure 9 report row-normalized confusion matrices for two representative baselines, CMERC and ECERC, and our method, SEDAL. As shown in Figure 8, SEDAL strengthens the diagonal entries for all six emotions on IEMOCAP. Compared with CMERC and ECERC, it improves recall for Happy (68.75%), Sad (80.41%), Neutral (75.26%), Angry (70.59%), Excited (79.26%), and Frustrated (72.18%). The most visible reductions occur in transition-sensitive confusions: Frustrated-to-Neutral drops to 11.29%, Angry-to-Frustrated drops to 24.71%, and Excited-to-Happy drops to 12.04%. These categories are difficult because their lexical content and acoustic intensity often overlap, so local utterance evidence alone can propagate the wrong neighboring emotion. SEDAL alleviates this by aligning each utterance to multimodal emotion anchors and then modeling how

the current anchor is reached from the previous turn under same-speaker or cross-speaker transition prototypes.

In Figure 9, the MELD results show a complementary pattern. SEDAL obtains stronger recall on Neutral (83.36%), Surprise (68.68%), Fear (24.00%), and Sadness (35.58%) than both baselines, and it reduces several majority-class absorption errors, such as Surprise-to-Neutral from 12.46% (CMERC) and 14.59% (ECERC) to 10.32% and Sadness-to-Neutral from 35.58% (CMERC) abd 34.62% (ECERC) to 32.21%. This suggests that speaker-aware transition anchors help the model distinguish a genuinely neutral response from an affective shift that follows from the previous speaker. At the same time, MELD remains challenging for rare or closely related emotions. Fear is still frequently predicted as Neutral or Surprise, and Disgust and Anger retain nontrivial confusion with Neutral and each other. These residual errors are consistent with the class imbalance and short utterance style of MELD, where several minority emotions are expressed through brief, context-dependent cues. The analysis therefore supports the role of SEDAL in reducing context-induced confusion while also showing that rare-emotion discrimination remains the main source of remaining errors.