SPAR: PREDICTIVE RELATION ROUTING FOR MULTIMODAL EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal emotion recognition relies on integrating heterogeneous affective cues, yet different modality combinations can yield inconsistent predictions for the same sample. The contribution of a modality depends on both the sample and the accompanying evidence, making predictive disagreement alone insufficient for determining which combinations should guide the final decision. To address this challenge, we introduce SPAR (Subset Prediction and Adaptive Routing), a framework that learns adaptive fusion through predictive relations between modality-subset predictions. These relations characterize differences in predictive distributions together with the modalities shared by or exclusive to the corresponding subsets. SPAR comprises two modules. The Modality Subset Prediction (MSP) module uses shared attention readers and a common classification head to produce emotion predictions for every nonempty subset of the available modalities, preserving unimodal and joint interpretations as fusion candidates. The Predictive Relation Routing (PRR) module encodes these relations over all pairs of subset predictions and integrates them with subset prediction descriptors through relation-aware attention, learning sample-dependent weights for aggregating subset logits. Both modules are jointly optimized through classification supervision on the subset predictions and the fused output. Experiments on MELD and the four- and six-class IEMOCAP settings show that SPAR achieves the highest reported score in eight of nine dataset–metric comparisons with discriminative baselines.

1 INTRODUCTION

Multimodal emotion recognition (MER) integrates linguistic content, vocal expression, and visual behavior to infer human emotional states (Poria et al., 2017). These modalities provide different kinds of information: words convey semantic content, while prosody and facial behavior help convey how that content is expressed (Baltrusaitis et al., 2019). Recent research investigates attention-based fusion and graph-based interaction modeling to integrate these heterogeneous signals (Tsai et al., 2019; Wu et al., 2025; Sun et al., 2026). To adapt this integration to individual samples, adaptive fusion methods account for variations in modality importance, highlighting the need to determine how strongly different sources of evidence should influence a prediction (Fang et al., 2025; Jiang et al., 2026).

The interpretation of an emotional cue can depend on the evidence accompanying it. In controlled psychological experiments, faces along a happiness–sadness continuum received different judgments when paired with happy or sad voices, with stronger vocal effects for ambiguous facial expressions (de Gelder & Vroomen, 2000). These findings show that an accompanying modality can change an affective judgment even when the original cue remains unchanged. For computational MER, this dependence must be considered alongside the possibility of noisy or conflicting inputs, which can complicate the integration of additional evidence (Wu et al., 2025). These observations motivate a fusion mechanism that considers a modality in relation to the accompanying evidence. The challenge is to exploit useful combinations of affective cues while limiting the influence of misleading information, with the contribution of each combination adapted to the current sample.

Recent approaches address this challenge by learning sample-dependent weights for modalities or experts (Moon et al., 2026). EMOE uses unimodal prediction errors to supervise sample-dependent modality weighting and employs unimodal distillation to preserve modality-specific predictive ability

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.

(Fang et al., 2025). I^2MoE trains experts for different interaction types through weakly supervised losses and learns to reweight their outputs (Xin et al., 2025). MiSTER-E dynamically combines the logits of separately supervised speech, text, and multimodal experts (Dutta et al., 2026). However, these weighting mechanisms do not explicitly associate pairwise differences between candidate predictions with the modality sets underlying those comparisons. Consequently, the relationship between predictive disagreement and the composition of the available evidence remains implicit when fusion weights are assigned.

Our formulation starts with a *modality-subset prediction*, defined as a prediction of the same utterance-level emotion label from a specified nonempty subset of modalities. We characterize *predictive relations* between subset predictions through differences in their predictive distributions together with the modalities shared by or exclusive to the corresponding subsets. These relations describe how predictions differ in the context of the evidence used to obtain them. Comparing text with text–audio examines the change associated with adding acoustic evidence, whereas comparing text–audio with text–visual contrasts two different sources of accompanying evidence. These comparisons may exhibit similar levels of predictive disagreement despite involving different changes in evidence. Disagreement magnitude alone therefore neither characterizes the evidence relation nor establishes which subset prediction should contribute more. Although a learned gate may infer such distinctions implicitly, explicitly encoding predictive and modality-set relations provides a structured basis for learning aggregation weights. This formulation calls for directly supervised predictions from the different combinations and a learned aggregation rule that evaluates each subset prediction through its relations with the others.

We address these requirements with SPAR (Subset Prediction and Adaptive Routing). Modality Subset Prediction (MSP) uses shared attention readers and a common classifier to preserve directly supervised predictions for every nonempty modality subset. Predictive Relation Routing (PRR) constructs a directed graph over these predictions, combining prediction descriptors with probability differences, distributional divergence, and shared or exclusive modalities. Relation-aware attention learns one sample-dependent scalar weight per subset; the weighted sum of subset logits determines the final distribution. Classification losses on subset predictions and the fused output jointly optimize the framework.

Our contributions are threefold:

- We develop MSP to retain directly supervised predictions for every nonempty subset of the available modalities through shared attention readers and a common classification head, making unimodal and joint interpretations available as fusion candidates.
- We propose PRR to learn the contributions of these subset predictions through explicit pairwise predictive and modality-set relations, connecting comparisons between evidence combinations to sample-dependent logit aggregation.
- We evaluate SPAR on MELD and the four- and six-class IEMOCAP settings, obtaining the highest reported scores in eight of nine dataset–metric comparisons with discriminative baselines.

2 Related Work

Modality combinations and subset prediction. Tensor fusion and its low-rank variants model interactions between modality representations (Zadeh et al., 2017; Liu et al., 2018). CSS combines low-rank polynomial fusion with gradient coordination across classification, unimodal regularization, and distillation objectives (Lyu et al., 2026). MISA separates modality-invariant and modality-specific representations (Hazarika et al., 2020), while MAG integrates nonverbal information into pretrained language representations (Rahman et al., 2020). MDF combines modality-aware disentanglement with adaptive modality weighting (Jian et al., 2026). DnR separates and refines modality-unique, pairwise-redundant, and synergistic representations for emotion recognition (Mai et al., 2026). DFG organizes unimodal, bimodal, and trimodal representations in a graph with input-dependent edge weights (Zadeh et al., 2018), while Multimodal Routing routes features from all seven nonempty combinations to label-associated concepts (Tsai et al., 2020). For MER with incomplete inputs, DQF uses modality-specific and cross-modal queries, applies auxiliary supervision to these combinations, and predicts from their concatenated representations (Miyoshi et al., 2026). HyperEF reconstructs

missing modality features with hypergraph-conditioned diffusion and performs uncertainty-aware evidence fusion (Qiu et al., 2026). Building on these precedents, SPAR's Modality Subset Prediction (MSP) module uses shared attention readers and a common classifier to retain directly supervised emotion predictions for every nonempty subset of the available modalities as aggregation candidates.

Adaptive modality and expert aggregation. EMOE supervises feature-conditioned modality weights using unimodal prediction errors and preserves modality-specific information through distillation (Fang et al., 2025). MMoE specializes interaction experts for redundancy, uniqueness, and synergy using unimodal prediction–label agreement patterns (Yu et al., 2024), whereas $\mathrm{I}^2\mathrm{MoE}$ uses weakly supervised interaction objectives and learns to reweight expert predictions (Xin et al., 2025). MiSTER-E directly supervises speech-only, text-only, and multimodal experts and adaptively combines their logits through a learned gate (Dutta et al., 2026). These methods provide precedents for prediction-informed learning and adaptive aggregation. SPAR focuses on explicitly connecting comparisons between candidate predictions to their underlying modality subsets.

Predictive relations and conflict-aware routing. Conversational models such as DialogueGCN encode relations between utterances and speakers (Ghosal et al., 2019). DMD learns directed distillation strengths from modality representations and logits to regulate knowledge transfer (Li et al., 2023). TiCAL uses pseudo unimodal labels and typicality to estimate consistency and guide learning (Yin et al., 2026). DCR performs contextual-bandit arbitration between fused and unimodal predictions and also evaluates an expanded seven-candidate action space (Yu et al., 2026). In SPAR, Predictive Relation Routing (PRR) encodes pairwise probability differences and distributional divergence together with the modalities shared by or exclusive to each subset pair. These comparisons describe predictive disagreement in the context of the evidence used to produce each prediction. Relation-aware attention integrates these relations with candidate predictive descriptors to learn sample-dependent weights for differentiable aggregation of subset logits. This makes evidence composition explicit when determining each candidate's contribution.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Let $\mathcal{M} = \{A, T, V\}$ denote audio, text, and vision, and let $\varnothing \neq \mathcal{A} \subseteq \mathcal{M}$ be the modalities available for an utterance. The observed inputs are $x_{\mathcal{A}} = \{x^m\}_{m\in\mathcal{A}}$, with an emotion label $y \in \{1, \ldots, K\}$. Given utterance-level supervision, the task is to learn a classifier F_θ that maps these inputs to an emotion distribution and a predicted label:

$$\widehat{p} = F_\theta(x_{\mathcal{A}}, \mathcal{A}) \in \Delta^{K-1}, \qquad \widehat{y} = \arg\max_{c\in\{1,\ldots,K\}} \widehat{p}_c, \tag{1}$$

where Δ^{K-1} denotes the probability simplex over K emotion classes. We use only the target utterance; dialogue context and speaker identity are not inputs to the fusion network.

SPAR constructs predictions from the nonempty modality subsets $\mathcal{S}(\mathcal{A}) = \{S : \varnothing \neq S \subseteq \mathcal{A}\}$, with $N = |\mathcal{S}(\mathcal{A})| = 2^{|\mathcal{A}|} - 1$. With complete inputs, these subsets are A, T, V, AT, AV, TV, ATV, where concatenated letters denote a modality set. A *modality subset* specifies the inputs used, whereas its *subset prediction*, represented by logits z_S and probabilities p_S, estimates the utterance's emotion from those inputs. Each subset prediction serves as a *fusion candidate*.

SPAR comprises two modules. Modality Subset Prediction (MSP) produces directly supervised predictions for all subsets in $\mathcal{S}(\mathcal{A})$. Predictive Relation Routing (PRR) compares these predictions through their predictive differences and modality-set relations, then learns sample-dependent weights for aggregating their logits. Figure 1 summarizes this computation.

3.2 MODALITY SUBSET PREDICTION

MSP projects modality features into a common space, summarizes each subset through shared attention readers, and applies a common classifier to obtain directly supervised subset predictions.

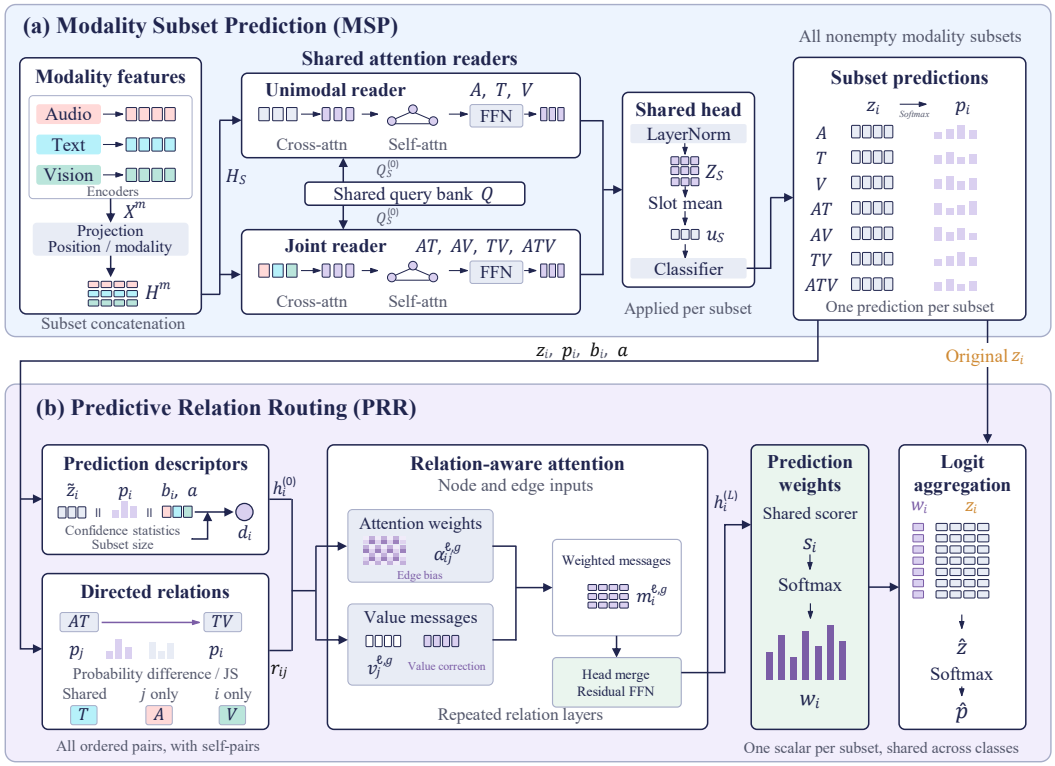


Figure 1: Overview of SPAR with complete audio, text, and vision inputs. (a) MSP uses shared attention readers and a common classifier to produce seven subset predictions. (b) PRR compares prediction descriptors through directed predictive and modality-set relations, then learns class-shared weights to aggregate subset logits. All predictions share utterance-label supervision; depicted dimensions and weights are schematic.

Modality feature preparation. For modality m, an encoder f_m transforms the input into a feature sequence $X^m = f_m(x^m) \in \mathbb{R}^{L_m \times d_m}$. A learned affine projection P_m maps these features to a common width d. Sinusoidal positional encodings π_t (Vaswani et al., 2017) and a learned modality embedding $e_m \in \mathbb{R}^d$ identify token positions and modality membership:

$$H_t^m = P_m(X_t^m) + \pi_t + e_m, \qquad t = 1, \ldots, L_m. \tag{2}$$

For subset S, we concatenate $\{H^m : m \in S\}$ along the sequence dimension to form H_S. This sequence supplies the information read by the attention reader. The modality sequences may have different lengths; reader masking and sequence-length correction are specified in Appendix A.

Shared attention readers. MSP summarizes each subset with *learnable query vectors* $Q \in \mathbb{R}^{R \times d}$. The slots attend to modality features to form sample-dependent representations, following attention-based pooling in Set Transformer and Perceiver (Lee et al., 2019; Jaegle et al., 2021).

The query bank initializes a unimodal reader $\mathcal{R}_{\text{uni}}$ shared across singleton subsets and a joint reader $\mathcal{R}_{\text{joint}}$ shared across multimodal subsets. With $\bar{e}_S = |S|^{-1} \sum_{m \in S} e_m$, modality identities enter the initial queries as

$$Q_S^{(0)} = \begin{cases} Q + \mathbf{1}_R e_m^\top, & S = \{m\}, \\ Q + QW_0 + \mathbf{1}_R \bar{e}_S^\top, & |S| > 1, \end{cases} \tag{3}$$

where $W_0 \in \mathbb{R}^{d \times d}$ is a learned transformation for joint-subset initialization. The query parameters are optimized during training. For a given subset, the initialized queries are sample-independent, while their subsequent reader states depend on the input features.

Each reader applies cross-attention to H_S, slot self-attention, and a feed-forward network. We normalize and average the output slots:

$$Z_S = \mathrm{LN}\Big(\mathcal{R}_{\tau(S)}(Q_S^{(0)}, H_S)\Big), \qquad u_S = \frac{1}{R}\sum_{r=1}^{R} Z_{S,r}, \tag{4}$$

where $\tau(S)$ selects the unimodal or joint reader. Both readers access the projected features of the selected modalities directly. The reader operator includes the masking and sequence-length correction detailed in Appendix A.

Shared classification head. A common affine head maps each subset representation into the same K-class emotion prediction space:

$$z_S = W_{\mathrm{cls}} u_S + c_{\mathrm{cls}}, \qquad p_S = \mathrm{softmax}(z_S). \tag{5}$$

Each p_S predicts the utterance's emotion using only the modalities in S. All subset predictions are supervised against the same utterance label y, as specified in Section 3.4. The resulting logits and probabilities provide the prediction-level inputs to PRR.

3.3 Predictive Relation Routing

PRR learns how strongly each subset prediction should contribute to the final decision by comparing it with the others. Its computation comprises four steps: *prediction descriptors* summarize individual predictions and their modality composition; *directed predictive relations* describe pairwise distributional differences together with shared and exclusive modalities; *relation-aware attention* integrates these node and edge descriptors; and *prediction weighting and aggregation* combines the subset logits using learned sample-dependent weights.

Each subset prediction forms a graph node, and its relations with other predictions form directed edges, including self-edges. We index the subsets as $S_1, \ldots, S_N$ and write $z_i = z_{S_i}$ and $p_i = p_{S_i}$.

Prediction descriptors. To represent each subset prediction as a graph node, we combine its class-specific outputs, confidence statistics, and modality composition. Let $a \in \{0,1\}^3$ indicate available modalities and $b_i \in \{0,1\}^3$ indicate membership in subset S_i. We center the logits as $\widetilde{z}_i = z_i - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top z_i$ and define

$$\begin{aligned} d_i &= \big[\widetilde{z}_i \,\|\, p_i \,\|\, H(p_i) \,\|\, \mu_i \,\|\, \rho_i \,\|\, b_i \,\|\, a \,\|\, |S_i|/3\big], \\ h_i^{(0)} &= \mathrm{GELU}(\mathrm{LN}(W_d d_i + c_d)). \end{aligned} \tag{6}$$

Here $\|$ denotes concatenation, $H(p) = -\sum_c p_c \log p_c$ is entropy, $\mu_i = p_{i,(1)} - p_{i,(2)}$ is the difference between the two largest probabilities, and $\rho_i = (K^{-1}\|\widetilde{z}_i\|_2^2 + \epsilon)^{1/2}$ is the centered logit root-mean-square, with $\epsilon = 10^{-12}$. The descriptor d_i has $2K + 10$ dimensions and is embedded into the initial node state $h_i^{(0)}$.

Directed predictive relations. To describe how two subset predictions differ in relation to their input composition, we construct a directed edge descriptor for every ordered pair. For receiving node i and sending node j, including $i = j$, define

$$\begin{aligned} r_{ij} = \big[&(p_j - p_i) \,\|\, \mathrm{JS}(p_i, p_j) \,\|\, (b_i \odot b_j) \\ &\|\, ((\mathbf{1} - b_i) \odot b_j) \,\|\, (b_i \odot (\mathbf{1} - b_j))\big], \end{aligned} \tag{7}$$

where $\odot$ denotes elementwise multiplication and

$$\mathrm{JS}(p_i, p_j) = \tfrac{1}{2}\,\mathrm{KL}(p_i \| \bar{p}_{ij}) + \tfrac{1}{2}\,\mathrm{KL}(p_j \| \bar{p}_{ij}), \qquad \bar{p}_{ij} = \tfrac{1}{2}(p_i + p_j). \tag{8}$$

The probability difference preserves the direction of class-specific changes, while Jensen–Shannon divergence summarizes distributional disagreement. The three masks identify shared, sender-only, and receiver-only modalities, respectively. Together, these components form a $K + 10$-dimensional descriptor that associates predictive differences with the evidence composition underlying each comparison.

For example, comparing T with AT describes the prediction change associated with including audio alongside text, whereas comparing AT with TV contrasts two different sources of accompanying evidence. Complete trimodal input therefore induces $7^2 = 49$ ordered pairs. Self-pairs have zero probability difference and divergence but retain their subset-intersection mask. These descriptors characterize relationships between model predictions rather than causal effects of changing the inputs.

Relation-aware attention. To assess each subset prediction using its relations with the others, PRR updates node states through relation-aware attention. At layer ℓ and head g, normalized node states give projected queries $q_i^{\ell,g}$, keys $k_j^{\ell,g}$, and values $v_j^{\ell,g}$. Following the use of pairwise relations in attention (Shaw et al., 2018), the predictive relations influence both attention scores and transmitted values:

$$\alpha_{ij}^{\ell,g} = \mathrm{softmax}_j\left(\frac{\langle q_i^{\ell,g}, k_j^{\ell,g}\rangle}{\sqrt{d_h}} + \beta_g^{\ell}(r_{ij})\right), \tag{9}$$

$$m_i^{\ell,g} = \sum_{j=1}^{N} \alpha_{ij}^{\ell,g}\left(v_j^{\ell,g} + E_g^{\ell}(r_{ij})\right). \tag{10}$$

Here d_h is the head width, β^{ℓ} is an MLP that returns one bias per head, and E^{ℓ} is an affine edge projection whose output is split across heads. Concatenated messages undergo an output projection, residual addition, and a pre-normalized feed-forward residual update:

$$\bar{h}_i^{\ell} = h_i^{\ell} + D\left(W_O^{\ell} \Big\|_g m_i^{\ell,g} + c_O^{\ell}\right), \tag{11}$$

$$h_i^{\ell+1} = \bar{h}_i^{\ell} + D\Big(\mathrm{FFN}^{\ell}(\mathrm{LN}(\bar{h}_i^{\ell}))\Big), \tag{12}$$

where D denotes dropout during training. The edge descriptors are computed once from the MSP predictions and reused across layers, with independently learned projections in each layer. The updated node states therefore combine information about individual subset predictions with their relations to the others.

Subset prediction weighting and aggregation. A shared linear scorer maps the updated node states to sample-dependent aggregation weights. After L relation layers,

$$s_i = w_o^{\top} h_i^{(L)} + c_o, \qquad w_i = \frac{\exp(s_i)}{\sum_{j=1}^{N}\exp(s_j)}, \qquad \widehat{z} = \sum_{i=1}^{N} w_i z_i, \quad \widehat{p} = \mathrm{softmax}(\widehat{z}). \tag{13}$$

Each subset prediction receives one scalar weight shared across emotion classes. The final distribution is obtained by applying softmax to the weighted sum of subset logits. All subset predictions are computed before aggregation, so the routing weights determine their contributions to the final decision.

3.4 Joint optimization

Building on the subset predictions produced by MSP and the fused output obtained through PRR, we jointly optimize the two modules with utterance-level classification supervision. For a training example $(x_{\mathcal{A}}, y)$ and available set $\mathcal{A}$, we minimize

$$\mathcal{L}(x_{\mathcal{A}}, y, \mathcal{A}) = \mathrm{CE}(\widehat{z}, y) + \frac{\lambda}{|\mathcal{S}(\mathcal{A})|}\sum_{S\in\mathcal{S}(\mathcal{A})} \mathrm{CE}(z_S, y). \tag{14}$$

Here CE denotes cross-entropy without class weights. The first term supervises the fused prediction; the auxiliary term averages classification losses over the available subsets, directly supervising each subset prediction regardless of its routing weight. Setting $\lambda = 1$ gives equal weight to the fused loss and the averaged subset loss.

Training samples modality availability as specified in Appendix B.2; only the supported encoders and subset predictors are evaluated. Full-input inference uses all seven subset predictions.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and baselines. We define a unified evaluation protocol for utterance-level multimodal emotion recognition on MELD (Poria et al., 2019) and IEMOCAP (Busso et al., 2008), using the text, audio, and video of the current utterance without neighboring utterances or speaker history. MELD contains seven emotion categories: neutral, surprise, fear, sadness, joy, disgust, and anger.

For IEMOCAP, we include both scripted and improvised dialogues. The dataset is divided into five sessions, where the first four sessions are used for training, while the last one is for evaluation. The six-class setting retains neutral, happy, sad, angry, frustrated, and excited as distinct categories. The four-class setting merges excited into happy, retains neutral, happy, sad, and angry, and excludes other labels.

Discriminative methods include DCR-utter (Yu et al., 2026), ALMT (Zhang et al., 2023), DMD (Li et al., 2023), PDCC-MER (Zhou et al., 2026), DQF (Miyoshi et al., 2026), and TiCAL (Yin et al., 2026). DCR-utter is our utterance-level reproduction of DCR. *Generative LLMs* include Emotion-LLaMA (Cheng et al., 2024), AffectGPT (Lian et al., 2025), and Nano-EmoX (Huang et al., 2026). Appendix F.3 details baseline selection and reproduction.

Implementation details. The modality encoders are RoBERTa-large (Facebook AI, n.d.) for text, Whisper-large-v3 (Radford et al., 2023) for audio, and CLIP ViT-B/16 (Radford et al., 2021) for vision. Feature extraction is detailed in Appendix B.2. MSP uses a hidden width of 256 and 16 query slots; both the unimodal and joint readers have four layers and four attention heads, with dropout 0.1. PRR has three layers, four heads, width 256, and dropout 0.2. A dataset-specific supervised warm start initializes the encoders and MSP, while PRR is initialized afresh with a zero-initialized final scorer. The encoders, MSP, and PRR are then jointly optimized.

We use AdamW (Loshchilov & Hutter, 2019) with peak learning rates of 10^{-6}, 10^{-5}, and 5×10^{-5} for the encoders, MSP, and PRR, respectively. Weight decay is 0.01 for linear and attention weight matrices and zero for other parameters. The effective batch size is 32, with 100 linear warmup updates and gradient clipping at a global norm of 1.0. Training runs for at most 40 continuation epochs with an early-stopping patience of ten. The objective uses unweighted cross-entropy and auxiliary coefficient $\lambda = 1$. Appendix B provides further implementation details.

Evaluation metrics. We report accuracy (ACC), weighted F1 (WF1), and macro-F1. Appendix E reports class recalls, whose unweighted mean is unweighted average recall (UAR). We additionally use emotion-wheel hit rate (HIT) to measure whether predicted emotions match the ground truth after semantic mapping. Details of HIT computation for SPAR are provided in Appendix D.

4.2 COMPARISON WITH THE STATE OF THE ART

Table 1 shows that SPAR leads eight of nine discriminative dataset–metric comparisons. On MELD, its ACC and WF1 exceed TiCAL by 1.58 and 0.92 percentage points, respectively, while macro-F1 exceeds DQF by 0.51 points. On IEMOCAP-4, SPAR improves over TiCAL by 1.29, 1.12, and 0.73 points in ACC, WF1, and macro-F1, respectively, showing gains under both support-weighted and equal-class evaluation. The advantage is narrower on IEMOCAP-6: ACC and WF1 improve over TiCAL by 0.49 and 0.43 points, whereas macro-F1 decreases by 0.15 points. Thus, the gains in overall classification do not extend uniformly to performance averaged equally across classes. The class-wise results in Appendix E further show that SPAR leads neutral and sad recall on IEMOCAP-6 but trails the best values for the other emotions.

For HIT, SPAR reaches 80.51 on IEMOCAP-4, exceeding the highest quoted generative score, AffectGPT's 60.54, by 19.97 points. Its MELD score of 56.45 is close to Nano-EmoX's 56.55. The IEMOCAP-4 result accompanies recalls above 77% for happy, sad, and angry, the emotions retained after neutral is excluded from HIT. MELD recalls remain low for fear (16.33%) and disgust (19.40%), revealing uneven class coverage despite the leading classification scores. This panel compares recognition after semantic mapping using published generative results; it does not assess open-ended emotion generation or explanation.

Table 1: Comparison with discriminative models (top) and generative models (bottom) on MELD and IEMOCAP. All scores are percentages. Best and second-best values within each column and panel are **bold** and underlined, respectively. * indicates results directly quoted from Table 2 of the Nano-EmoX paper. Class-wise classification results are provided in Appendix E.

Comparison with discriminative models

	MELD			IEMOCAP					
Method	ACC-7	WF1	Macro-F1	ACC-6	WF1	Macro-F1	ACC-4	WF1	Macro-F1
DMD	60.73	59.76	41.66	60.11	59.50	57.44	77.03	77.24	77.00
ALMT	61.97	59.55	36.67	58.26	58.48	57.78	76.23	76.23	76.04
DCR-utter	60.92	60.97	42.36	61.59	61.49	59.76	80.02	80.13	80.27
PDCC-MER	63.70	62.33	43.47	64.06	64.06	62.91	80.34	80.33	79.73
DQF	64.39	63.52	49.27	63.75	64.13	63.10	79.94	80.01	79.43
TiCAL	65.16	65.20	45.29	65.54	65.57	**64.35**	80.66	80.97	81.21
SPAR (Ours)	**66.74**	**66.12**	**49.78**	**66.03**	**66.00**	64.20	**81.95**	**82.09**	**81.94**

Comparison with generative models

	MELD	IEMOCAP-4
Method	HIT	HIT
Emotion-LLaMA*	46.76	55.47
AffectGPT*	55.65	60.54
Nano-EmoX*	**56.55**	60.12
SPAR (Ours)	56.45	**80.51**

4.3 Ablation Study

Table 2 evaluates three components on MELD. For *MSP subset predictions*, SingleSubset uses only the complete ATV prediction without PRR; NoBimodal excludes bimodal candidates but retains their auxiliary supervision. For *PRR aggregation and relations*, Uniform assigns weight $1/N$ to each candidate. **MLPGate** replaces PRR with a shared MLP that independently scores each candidate using the same node descriptors, followed by softmax normalization to obtain fusion weights. This variant assesses the benefit of interactions among candidates beyond candidate-wise gating. **NodeAttn** uses node-only multi-head self-attention, removing both edge-conditioned attention biases and edge-dependent message terms to examine the contribution of explicit relational information beyond standard attention. NoPredEdges zeroes predictive edge components, retaining modality-set relations and node descriptors. For *subset auxiliary supervision*, NoAux removes the auxiliary loss during joint optimization. All variants share a supervised warm start; Appendix F.1 details their continuation-stage configurations.

Table 2: Ablation results on MELD (%). Best and second-best values in each column are **bold** and underlined, respectively.

Analysis	Configuration	ACC-7	WF1	Macro-F1
Reference	Full SPAR	**66.74**	**66.12**	**49.78**
MSP subset predictions	SingleSubset	63.82	62.13	42.66
	NoBimodal	65.59	64.01	46.65
PRR aggregation and relations	Uniform	63.12	62.26	43.81
	MLPGate	64.62	62.43	43.56
	NodeAttn	65.82	63.10	44.13
	NoPredEdges	66.17	64.10	46.53
Subset auxiliary supervision	NoAux	62.89	62.40	44.93

Full SPAR leads all three metrics in Table 2. SingleSubset loses 3.99 WF1 points and 7.12 macro-F1 points. NoBimodal retains bimodal auxiliary supervision but loses 2.11 and 3.13 points in these metrics, supporting the use of bimodal candidates in aggregation.

Uniform loses 3.86 WF1 points and 5.97 macro-F1 points. MLPGate recovers 1.50 ACC points over Uniform but only 0.17 WF1 points, while macro-F1 falls by 0.25 points. NodeAttn improves on MLPGate across all three metrics but trails Full by 0.92 ACC points, 3.02 WF1 points, and 5.65 macro-F1 points. These comparisons support candidate interactions and explicit relational information beyond independent gating and standard attention. NoPredEdges achieves the second-best ACC and WF1, yet loses 2.02 WF1 points and 3.25 macro-F1 points relative to Full, supporting predictive edge features alongside the retained predictive node inputs and modality-set relations.

NoAux yields the lowest ACC (62.89), losing 3.72 WF1 points and 4.85 macro-F1 points. These drops support subset supervision during joint optimization beyond the shared supervised warm start.

4.4 Case Analysis

We compare Full SPAR (Full) and uniform aggregation (Uniform) on MELD using a fixed reference MSP. The number of distinct unimodal top-1 labels is

$$B = \left| \left\{ \arg \max_{c \in \{1,\dots,K\}} p^{\mathrm{ref}}_{\{m\},c} : m \in \mathcal{M} \right\} \right|, \tag{15}$$

Here $p^{\mathrm{ref}}_{\{m\},c}$ denotes a reference MSP probability. $B = 1, 2, 3$ mean unanimous, two agreeing, and three distinct predictions. Both methods share groups defined without ground-truth labels.

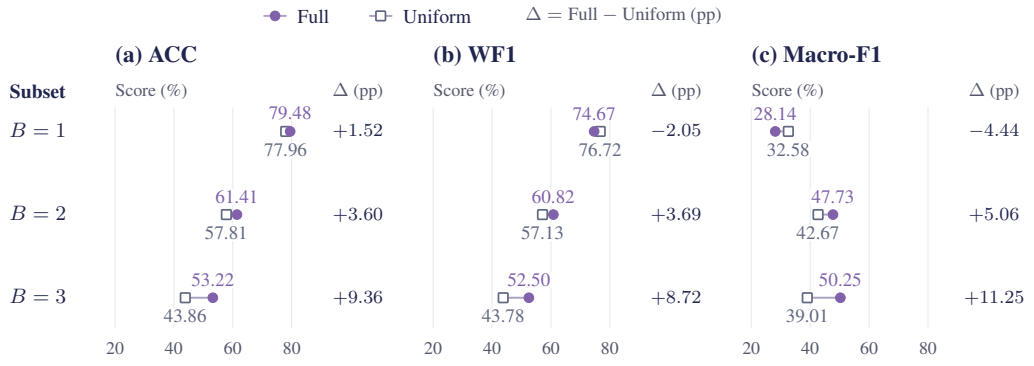


Figure 2: Full and Uniform on MELD: scores (%) and differences Δ (pp).

Figure 2 reports ACC, WF1, and macro-F1 for the three groups. For $B = 1$, Full improves ACC by 1.52 percentage points but lowers WF1 and macro-F1 by 2.05 and 4.44 points, respectively, revealing a trade-off between overall accuracy and class-wise F1 under unanimous predictions. For $B = 2$, Full improves all three metrics, with gains of 3.60, 3.69, and 5.06 points in ACC, WF1, and macro-F1, respectively. The largest gains occur for $B = 3$: Full reaches 53.22 ACC, 52.50 WF1, and 50.25 macro-F1, exceeding Uniform by 9.36, 8.72, and 11.25 points. These results suggest greater benefits from adaptive aggregation when unimodal predictions differ, while its effect under unanimous predictions remains metric-dependent. Predictive disagreement does not itself establish semantic conflict between the input modalities.

5 Conclusion

We presented SPAR, combining directly supervised modality-subset predictions through MSP with sample-dependent logit aggregation through PRR using predictive differences and modality-set relations. SPAR leads eight of nine dataset–metric comparisons with discriminative baselines across MELD and both IEMOCAP settings. MELD ablations support subset diversity, candidate interactions, explicit relational information, and continued auxiliary supervision. The case analysis shows the largest gains over uniform aggregation when all three unimodal predictions differ. Future work will examine fully separated model selection and testing, missing or degraded modalities, and subset selection for larger modality inventories.

AI USE STATEMENT

We used AI tools for literature searches, manuscript drafting and polishing, LaTeX preparation, assistance with reproducing selected experiments, and feedback on ablation design and interpretation of experimental results. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The main text specifies the model design and training objective (Section 3), datasets, evaluation protocol, implementation settings, and ablation configurations (Section 4). Appendix A derives the sequence-length correction and details the attention readers; Appendix B provides the architecture and input-processing settings. Appendix D defines HIT scoring, and Appendix E reports class-wise results. Appendix F details the ablation configurations and case-analysis grouping. These descriptions specify the computations and evaluation conventions underlying the reported experiments.

REFERENCES

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019.

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.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Beatrice de Gelder and Jean Vroomen. The perception of emotions by ear and by eye. *Cognition and Emotion*, 14(3):289–311, 2000.

Soumya Dutta, Smruthi Balaji, and Sriram Ganapathy. A mixture-of-experts model for multimodal emotion recognition in conversations. *Computer Speech & Language*, 100:101965, 2026.

Facebook AI. RoBERTa-large: Model card. Official model documentation, n.d.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. EMOE: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14314–14324, 2025.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 154–164. Association for Computational Linguistics, 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*, pp. 1122–1131. Association for Computing Machinery, 2020.

Jiahao Huang, Fengyan Lin, Xuechao Yang, Chen Feng, Kexin Zhu, Xu Yang, and Zhide Chen. Nano-EmoX: Unifying multimodal emotional intelligence from perception to empathy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22986–22997, 2026.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021.

Zhongquan Jian, Wenhan Lv, Yanhao Chen, Guanran Luo, Wentao Qiu, Shaopan Wang, Bingbing Hu, and Qingqiang Wu. MDF: A modality-aware disentanglement and fusion framework for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31292–31300, 2026.

Haoyu Jiang, Xiaoliang Chen, Duoqian Miao, Xiaolin Qin, Xianyong Li, and Yajun Du. CICA: Coupling confidence-aware pretraining with confidence-informed attention for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37842–37851, 2026.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

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

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36993–37014. PMLR, 2025.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 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)*, pp. 2247–2256. Association for Computational Linguistics, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xiaosen Lyu, Jiayu Xiong, Yuren Chen, Wanlong Wang, Xiaoqing Dai, and Jing Wang. Cross-space synergy: A unified framework for multimodal emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24226–24234. AAAI Press, 2026.

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*, pp. 2700–2709, 2026.

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

A-Seong Moon, Haesung Kim, Ye-Chan Park, and Jaesung Lee. A survey on multimodal emotion recognition: Methods, datasets, and future directions. *Computers, Materials & Continua*, 87(2):1, 2026.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

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*, pp. 527–536. Association for Computational Linguistics, 2019.

Xihang Qiu, Yuhao Fang, Qing Zhou, Bin Zhai, Jialong Hong, Wanpeng Zhang, Yao Lu, Ye Zhang, and Chun Li. Beyond missing modalities: Hypergraph conditioned diffusion for uncertainty-aware multimodal emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22953–22963, 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 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

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

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468. Association for Computational Linguistics, 2018.

Jiangfeng Sun, Sihao He, Zhonghong Ou, and Meina Song. Structures meet semantics: Multimodal fusion via graph contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25691–25699, 2026.

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

Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1823–1833. Association for Computational Linguistics, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017.

ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6257–6274. Association for Computational Linguistics, 2025.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L. Ballard, Tianlong Chen, and Qi Long. $\mathrm{I}^2\mathrm{MoE}$: Interpretable multimodal interaction-aware mixture-of-experts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68870–68888. PMLR, 2025.

Wen Yin, Siyu Zhan, Cencen Liu, Xin Hu, Guiduo Duan, Xiurui Xie, Yuan-Fang Li, and Tao He. TiCAL: Typicality-based consistency-aware learning for multimodal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17948–17956. AAAI Press, 2026.

Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. MMoE: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10006–10030. Association for Computational Linguistics, 2024.

Yangchen Yu, Qian Chen, Jia Li, Zhenzhen Hu, Jinpeng Hu, Lizi Liao, Erik Cambria, and Richang Hong. To fuse or to drop? dual-path learning for resolving modality conflicts in multimodal emotion recognition. arXiv:2605.04877, 2026.

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

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

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 756–767. Association for Computational Linguistics, 2023.

Xin Zhou, Xiang Shen, Dezhi Han, Yangshuyi Xu, Xiaoqi Xin, and Kuan-Ching Li. PDCC-MER: Progressive denoising and calibrated consensus learning for multimodal emotion recognition. *Knowledge-Based Systems*, 352:116975, 2026.

A ATTENTION READERS AND SEQUENCE-LENGTH CORRECTION

Each reader contains four blocks with independent parameters across depth. Let b_S denote the additive cross-attention bias over positions in the subset memory H_S, specified below for valid tokens and padding. For query slots $Q^{(\ell)}$ and fixed subset memory H_S, a block computes

$$Q' = Q^{(\ell)} + D\big(\mathrm{MHA}_{\mathrm{cross}}(\mathrm{LN}_q(Q^{(\ell)}), \mathrm{LN}_h(H_S), \mathrm{LN}_h(H_S); b_S)\big), \tag{16}$$

$$Q'' = Q' + D\big(\mathrm{MHA}_{\mathrm{self}}(\mathrm{LN}_s(Q'), \mathrm{LN}_s(Q'), \mathrm{LN}_s(Q'))\big), \tag{17}$$

$$Q^{(\ell+1)} = Q'' + D\big(\mathrm{FFN}(\mathrm{LN}_f(Q''))\big). \tag{18}$$

The notation $\mathrm{MHA}(Q, K, V; b)$ specifies queries, keys, values, and an additive bias broadcast over query positions and attention heads. All normalization and attention modules are distinct within a block. The feed-forward map is $256 \to 1024 \to 256$, with GELU and dropout between its linear layers. Both attention modules use four heads; attention dropout and residual dropout are 0.1. The output normalization in Eq. (4) is shared across subsets. Learned modality identities enter both the memories and the initial query slots. The positional encoding is recomputed within each modality sequence before memories are concatenated.

Masking and sequence-length correction. The reader operator in Eq. (4) includes an additive bias in cross-attention. Each valid token from modality m receives a bias of $-\log L_m$, while padded positions receive $-\infty$. The bias is broadcast over query slots and attention heads. This correction compensates for differences in token count: when all valid tokens have equal attention scores, each included modality receives equal total attention mass. Content-dependent attention scores can still assign different weights to the modalities.

For a valid position t in modality m, write its unnormalized attention score as $a_{m,t}$. The length-adjusted mass assigned to modality m is

$$A_m = \frac{L_m^{-1} \sum_{t=1}^{L_m} \exp(a_{m,t})}{\sum_{n \in S} L_n^{-1} \sum_{t=1}^{L_n} \exp(a_{n,t})}. \tag{19}$$

For a singleton subset the bias is constant over valid positions and therefore cancels in the softmax. Padding values are zeroed before projection, and padded positions are masked in attention. This correction addresses token-count imbalance in the reader; it does not establish that the modalities contain equal information or equal-quality evidence.

B IMPLEMENTATION AND TRAINING CONFIGURATION

B.1 ARCHITECTURE

SPAR is evaluated with $K = 7$ classes on MELD and with $K = 4$ or $K = 6$ classes on IEMOCAP. Table 3 specifies the shared architecture, whose classifier and prediction descriptors adapt to the number of classes. The node-feature dimension is $2K + 10$, giving 24, 18, and 22 features for the seven-, four-, and six-class tasks, respectively; the edge-feature dimension is $K + 10$, giving 17, 14, and 16 features. Natural logarithms are used throughout. Jensen–Shannon divergence is computed with log-softmax and log-sum-exp for numerical stability. Entropy uses a probability floor of 10^{-12} inside the logarithm.

Table 3: Architecture across the seven-, four-, and six-class tasks. Parameter counts exclude the modality encoders.

Component	Configuration
Encoder feature widths	Audio 1280; text 1024; vision 768
MSP width / slots	$d = 256$; $R = 16$
Unimodal / joint readers	Four layers each; four heads; distinct reader stacks
Reader FFN / dropout	$256 \to 1024 \to 256$; 0.1
Subset classifier	Shared affine $256 \to K$
PRR embedding	Affine $(2K + 10) \to 256$, LayerNorm, GELU
Relation network	Three layers; four heads; width 256; head width 64
Edge attention bias	Affine $(K + 10) \to 64$, GELU, affine $64 \to 4$
Edge value	Affine $(K + 10) \to 256$, then split into four heads
PRR FFN / dropout	$256 \to 1024 \to 256$; 0.2
Final scorer	Shared affine $256 \to 1$, initialized to zero
MSP parameters	$9{,}289{,}728 + 257K$
PRR parameters	$2{,}384{,}205 + 1{,}472K$
Total task parameters	$11{,}673{,}933 + 1{,}729K$

The PRR dropout is applied inside the feed-forward map and on the residual branches. It is not applied to the relation-attention probabilities. Candidate order in the implementation is A, T, V, AT, AV, TV, ATV, with bit masks $1, 2, 4, 3, 5, 6, 7$, respectively. This order is an indexing convention, not a positional encoding of candidates. Parameters of the node embedding, edge maps, and scorer are shared across candidate indices.

B.2 Inputs and encoders

The implementation uses the RoBERTa-large text encoder, the Whisper-large-v3 audio encoder, and the CLIP ViT-B/16 vision encoder. Whisper and CLIP are described in their final conference papers (Radford et al., 2023; 2021); the exact checkpoint identifiers are implementation choices. The text checkpoint and the later Whisper-v3 release are identified by their official model records.[1]

Text. Only the verbatim target utterance is tokenized. Right truncation limits the sequence to 512 tokens, retaining the tokenizer's special tokens. All valid last-layer token states are used, without a pooler. No neighboring utterance, dialogue identifier, or speaker identifier is supplied.

Audio. Audio is converted to mono at 16 kHz; stereo channels are averaged, and other channel layouts use the decoder's mono downmix. The retained preprocessing metadata specifies windows of at most 30 seconds. Each window becomes a 128×3000 log-mel array, with padding where needed. From the encoder output, the first $\min(1500, \lceil n/320 \rceil)$ positions are retained for a window with n waveform samples. Valid positions from successive windows are concatenated. The speech decoder is not used.

Vision. Frames are selected using the retained preprocessing metadata's frame-index lists. Each frame is converted to RGB, resized with bicubic interpolation so its longest side is 224 pixels, and centered in a 224×224 canvas filled with the rounded CLIP mean color. CLIP mean and standard deviation normalization is then applied. The visual sequence consists of each frame's last-layer class-token state before CLIP's pooling LayerNorm, arranged in the specified frame order.

Modality availability and baseline settings. During training, complete ATV inputs are sampled with probability $1/2$ and each other nonempty modality set with probability $1/12$; evaluation uses complete ATV inputs. The model also produces outputs for fewer modalities, but this architectural property does not establish empirical robustness to missing modalities. The reproduced discriminative

[1] Official implementation records: https://huggingface.co/FacebookAI/roberta-large and https://huggingface.co/openai/whisper-large-v3.

baselines retain their method-specific optimizers, learning rates, batch sizes, and encoder fine-tuning settings under the same data and selection protocol.

C SCALABILITY

The present trimodal formulation uses all nonempty subsets. For future extensions to more modalities, the subset family could be restricted to pairwise modality combinations and the full available-modality subset, with modality descriptors adapted to the enlarged modality inventory. For M available modalities, the number of distinct subsets would be $N' \leq \binom{M}{2} + 1$. The number of subset predictions would therefore grow quadratically with the number of available modalities, while relation attention would still compare $(N')^2$ ordered pairs, whose count grows quartically with M. MSP and PRR could operate on this selected family using the same prediction and routing principles, providing a route toward greater scalability.

D HIT EVALUATION

For each utterance, SPAR's final fused prediction is obtained by argmax and converted into a singleton emotion set. We apply the official AffectGPT/MER-UniBench (Lian et al., 2025) normalization, synonym mappings, and five level-1 emotion wheels. No top-k expansion or language-model-based label extraction is used.

Let $G_i^{(w)}$ and $P_i^{(w)}$ denote the mapped ground-truth and prediction sets under wheel w. HIT is calculated as

$$\text{HIT} = \frac{1}{5}\sum_{w=1}^{5}\frac{1}{|\mathcal{I}_w|}\sum_{i\in\mathcal{I}_w}\mathbf{1}\left[G_i^{(w)} \cap P_i^{(w)} \neq \varnothing\right], \tag{20}$$

where $\mathcal{I}_w$ contains samples with nonempty mapped ground truths. Each retained mapped ground truth is a singleton, so the indicator equals the intersection size used by the official scorer for these tasks. Unmapped ground truths, including neutral labels, are excluded; empty predictions remain misses.

E CLASS-WISE CLASSIFICATION RESULTS

Tables 4–6 supplement Table 1 with class-wise accuracy and the corresponding overall ACC, WF1, and macro-F1. For each emotion, class-wise accuracy is the percentage of utterances with that ground-truth label that are classified correctly, i.e., class recall. All entries are percentages rounded to two decimal places. Best and second-best values in each column are **bold** and underlined, respectively; ties receive the same marking.

Table 4: Class-wise accuracy (recall) and overall metrics on MELD-7 (%).

Method	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	ACC-7	WF1	Macro-F1
DMD (Li et al., 2023)	79.76	57.35	14.29	23.56	54.86	19.40	38.12	60.73	59.76	41.66
ALMT (Zhang et al., 2023)	79.36	65.23	0.00	13.94	59.10	0.00	49.27	61.97	59.55	36.67
DCR-utter (Yu et al., 2026)	74.96	68.82	4.08	31.25	56.11	34.33	40.18	60.92	60.97	42.36
PDCC-MER (Zhou et al., 2026)	**85.44**	49.46	**22.45**	20.19	62.34	22.39	37.83	63.70	62.33	43.47
DQF (Miyoshi et al., 2026)	82.00	49.82	16.33	22.12	51.62	**46.27**	**63.05**	64.39	63.52	49.27
TiCAL (Yin et al., 2026)	80.16	**73.48**	4.08	33.65	60.10	37.31	42.82	65.16	65.20	45.29
SPAR (Ours)	78.56	63.44	16.33	**34.62**	**68.58**	19.40	60.12	**66.74**	**66.12**	**49.78**

Table 5: Class-wise accuracy (recall) and overall metrics on IEMOCAP-6 (%).

Method	Neutral	Happy	Sad	Angry	Frustrated	Excited	ACC-6	WF1	Macro-F1
DMD (Li et al., 2023)	70.05	22.38	63.67	60.59	58.27	**64.55**	60.11	59.50	57.44
ALMT (Zhang et al., 2023)	63.80	48.95	65.31	61.76	55.91	50.84	58.26	58.48	57.78
DCR-utter (Yu et al., 2026)	59.90	33.57	66.12	65.29	**69.29**	61.54	61.59	61.49	59.76
PDCC-MER (Zhou et al., 2026)	71.88	47.55	65.71	**75.88**	66.40	50.84	64.06	64.06	62.91
DQF (Miyoshi et al., 2026)	69.01	**53.15**	70.61	67.06	60.63	58.53	63.75	64.13	63.10
TiCAL (Yin et al., 2026)	70.57	45.45	71.02	70.59	65.35	61.54	65.54	65.57	**64.35**
SPAR (Ours)	**75.26**	46.85	**77.55**	60.59	60.89	63.55	**66.03**	**66.00**	64.20

Table 6: Class-wise accuracy (recall) and overall metrics on IEMOCAP-4 (%).

Method	Neutral	Happy	Sad	Angry	ACC-4	WF1	Macro-F1
DMD (Li et al., 2023)	74.74	75.79	77.55	84.71	77.03	77.24	77.00
ALMT (Zhang et al., 2023)	80.47	68.78	**82.86**	76.47	76.23	76.23	76.04
DCR-utter (Yu et al., 2026)	79.17	78.05	81.63	84.71	80.02	80.13	80.27
PDCC-MER (Zhou et al., 2026)	80.99	**86.43**	69.39	78.82	80.34	80.33	79.73
DQF (Miyoshi et al., 2026)	78.65	83.48	74.69	81.18	79.94	80.01	79.43
TiCAL (Yin et al., 2026)	84.90	79.41	67.35	**93.53**	80.66	80.97	81.21
SPAR (Ours)	**85.16**	77.60	80.00	88.24	**81.95**	**82.09**	**81.94**

F Supplementary Experimental Details

F.1 Ablation Configurations

Table 2 compares MELD continuation-stage configurations using a common supervised warm start and Section 4.1's settings. SingleSubset trains and evaluates with complete ATV inputs, using only the trimodal prediction without PRR. This system-level comparison changes training and aggregation; other variants retain modality-availability sampling. NoBimodal retains all subset auxiliary supervision but routes only singleton and trimodal candidates, or the two singletons when two modalities are available. Uniform assigns weight $1/N$ to each candidate. MLPGate retains the node embedding and scalar scorer, independently scoring candidates through three residual FFN blocks shared across candidates, with pre-LayerNorm, model width 256, and FFN hidden width 1550. Softmax and logit aggregation follow Eq. (13); only normalization couples its candidate weights. NodeAttn removes $\beta_g^{\ell}(r_{ij})$ from Eq. (9) and $E_g^{\ell}(r_{ij})$ from Eq. (10), retaining three self-attention layers with FFN hidden width 1036. Both configurations maintain parameter counts comparable to full PRR, with all added parameters participating in the forward computation. NoPredEdges zeroes only the probability-difference and Jensen–Shannon components in Eq. (7), retaining the nine modality-set components. NodeAttn and NoPredEdges retain all predictive and modality-composition node features. NoAux removes the auxiliary term in Eq. (14) during joint optimization while retaining fused-output supervision. Section 4.3 uses Table 2's two-decimal scores to compute differences.

F.2 Case-Analysis Grouping and Metrics

The reference MSP in Eq. (15) defines one shared partition using only the singleton subsets; bimodal and trimodal predictions do not enter the grouping rule. Ground-truth labels are used only for the within-group metrics. WF1 weights class F1 scores by their support within the group, whereas macro-F1 gives each emotion class equal weight. These aggregate comparisons do not establish statistical significance or a causal effect of predictive disagreement.

F.3 BASELINE SELECTION AND REPRODUCTION

We select the baselines in Table 1 for compatibility with utterance-level text–audio–visual inputs, relevance to multimodal fusion, and coverage of complementary mechanisms. Methods developed for related sentiment tasks are included when their mechanisms can be adapted to categorical emotion recognition.

The discriminative group covers language-guided fusion (ALMT; Zhang et al., 2023), representation decomposition and cross-modal distillation (DMD; Li et al., 2023), progressive denoising and reliability-calibrated routing (PDCC-MER; Zhou et al., 2026), query-based unimodal and cross-modal learning (DQF; Miyoshi et al., 2026), consistency-aware learning (TiCAL; Yin et al., 2026), and arbitration between fused and unimodal predictions (DCR-utter; Yu et al., 2026). These choices examine alternatives to learning from relations among subset predictions. The generative group covers instruction-tuned emotion recognition and reasoning (Emotion-LLaMA; Cheng et al., 2024), descriptive emotion understanding (AffectGPT; Lian et al., 2025), and compact multitask affective modeling (Nano-EmoX). Their HIT results are quoted from Table 2 of Nano-EmoX (Huang et al., 2026), as marked in Table 1.

To ensure fair comparison and reproducibility, we base the discriminative implementations on the original architectures and reported hyperparameters, using official or public codebases where available and paper-based reimplementations otherwise. Necessary input and classification adaptations follow our evaluation protocol; DCR-utter uses only the current utterance.