Where Emotions Peak: Hotspot-Centric Multimodal Emotion Recognition in Conversations

Anonymous ACL submission

Abstract

Multimodal Emotion Recognition in Conversations (ERC) typically represents each modality as a single, time-aggregated utterance-level feature. This coarse-grained aggregation discards the fine-grained evidence that is often decisive for emotion discrimination—the transient prosodic burst distinguishing excitement from happiness, or the fleeting micro-expression separating composure from suppressed frustration. We refer to such localized, high-intensity segments as *emotion hotspots* and argue that ERC benefits from explicitly modeling emotion at multiple granularities rather than collapsing all temporal positions into a single vector. Crucially, fine-grained hotspots and global context are mutually enriching—each supplying what the other lacks—and must be coherently integrated across modalities. We therefore propose a granularity-aware framework in three stages: (i) modality-tailored hotspot extraction localizes fine-grained emotional evidence; (ii) *Hotspot-Gated Fusion* (HGF) adaptively interpolates between hotspot and global features via learned per-token gates, yielding multi-granularity unimodal representations; (iii) a routed *Mixture-of-Aligners* (MoA) flexibly aligns these enriched representations through utterance-level expert routing, complemented by a graph pathway encoding conversational structure. On IEMOCAP and CMU-MOSEI, the framework achieves competitive results, with ablations confirming that each granularity stage contributes gains.

1 Introduction

Emotion Recognition in Conversations (ERC) (Fu et al., 2023; Wu et al., 2025) aims to identify the emotion of each utterance in a dialogue from a predefined set of categories. Despite progress in sequence modeling (Majumder et al., 2019), graph-based reasoning (Ghosal et al., 2019), and multimodal fusion (Liu et al., 2020; Yu et al., 2024; Khan et al., 2025), existing methods share a common practice of relying on time-aggregated, per-modality representations at the utterance level—textual embeddings and pooled acoustic or visual descriptors—which we refer to as *global features*. Treating all temporal positions equally implicitly assumes that emotional cues are distributed uniformly across an utterance—an assumption rarely met in practice (Scherer, 2009; Mirsamadi et al., 2017; Wagner et al., 2023). Emotions instead tend to surface in short, high-intensity bursts—an affect-laden phrase, a pitch spike, a fleeting facial expression—that we term *emotion hotspots*. Both affective science and prosodic research support this view, showing that facial affect is conveyed through transient muscle activations (Ekman and Friesen, 1978) and that vocal cues cluster around local pitch and energy peaks (Madanian et al., 2023; Yan et al., 2013; Ekman, 2003). The central question, then, is how to move beyond coarse utterance-level aggregation and exploit emotion signals at their native, fine-grained temporal scale.

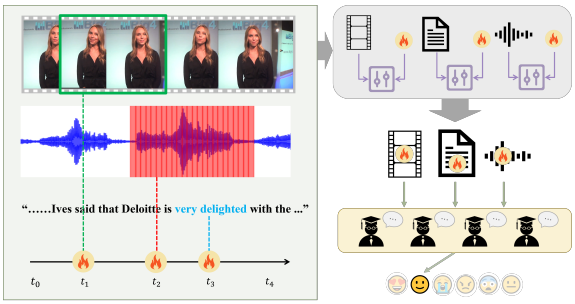

Figure 1: Overview of the granularity-aware framework: modality-specific hotspot extraction localizes fine-grained emotional evidence, Hotspot-Gated Fusion integrates it with global context, and MoA performs cross-modal re-alignment.

Operating at this finer granularity yields richer discriminative representations than utterance-level aggregation alone. Global pooling treats all temporal positions equally, inevitably averaging away

the sparse but decisive emotional peaks (Mirsamadi et al., 2017; Wang et al., 2016); by contrast, hotspot-level features preserve the precise acoustic, lexical, and visual cues at which emotions are most saliently expressed. For instance, disambiguating *excited* from *happy* hinges on a brief prosodic surge that mean pooling dilutes. Crucially, fine-grained hotspots and global context are mutually enriching: hotspots supply the precise emotional evidence that global pooling discards, while the global utterance scaffold disambiguates inherently ambiguous local cues—a pitch spike in isolation may signal anger or surprise, but its meaning crystallizes given the surrounding context (Liu et al., 2020; Barrett et al., 2019). Bridging these two granularity levels within each modality, and coherently integrating the enriched representations across modalities, remains an open challenge (Tsai et al., 2019; Jiang et al., 2025; Li et al., 2026).

To this end, we propose a granularity-aware framework that integrates hotspot extraction and local–global fusion as an organic whole, enabling the two granularity levels to mutually enrich each other rather than being treated as separable processing steps. Figure 1 provides an overview. Concretely, the framework operates in three progressive stages. First, we extract per-modality hotspots through strategies tailored to each signal, thereby localizing the fine-grained emotional evidence that coarse aggregation discards. Second, *Hotspot-Gated Fusion* (HGF) performs token-wise gated interpolation between the extracted hotspots and global utterance context, yielding multi-granularity unimodal representations that retain both precise local cues and the broader scaffold for disambiguation. Third, a routed *Mixture-of-Aligners* (MoA) flexibly re-aligns the enriched cross-modal representations through utterance-level expert routing, while a complementary graph pathway encodes conversational structure; the two streams are combined for final prediction. **Our main contributions are summarized as follows:**

- We design modality-specific **emotion hotspot extraction** strategies that localize the most emotionally discriminative segments within each modality, surfacing the fine-grained cues that global pooling averages away.
- We introduce **Hotspot-Gated Fusion**, a token-wise gated interpolation mechanism that adaptively blends fine-grained hotspot evidence with global utterance context, preserving both localized emotional peaks and the contextual scaffold needed for disambiguation.
- We couple HGF with a routed **Mixture-of-Aligners** for cross-modal re-alignment and a conversational graph pathway for dialogue structure; experiments on IEMOCAP and CMU-MOSEI show state-of-the-art results, with ablations confirming that each granularity stage contributes complementary gains.

2 Related Work

Our framework touches on four lines of research: ERC architectures that model conversational context, multimodal fusion strategies that combine heterogeneous signals, cross-modal alignment methods that resolve inter-modal discrepancies, and salient region detection techniques that localize emotionally informative segments.

2.1 Emotion Recognition in Conversations

ERC has evolved from recurrent sequence models (Majumder et al., 2019; Poria et al., 2017; Wang et al., 2022) to graph neural networks that capture long-range speaker-level and contextual dependencies (Wei et al., 2019; Hu et al., 2021; Joshi et al., 2022). CORECT (Nguyen et al., 2023) enriches relational temporal graphs with auxiliary cross-modality interaction, and M3NET (Chen et al., 2023) applies multi-frequency spectral analysis to GNN-based ERC. Beyond graph methods, COSMIC (Ghosal et al., 2020) injects commonsense knowledge to strengthen contextual reasoning. More recently, large language models have been applied to ERC: InstructERC (Lei et al., 2023) reformulates the task via retrieval-augmented instruction tuning, and DialogueLLM (Zhang et al., 2025) fine-tunes LLMs with emotion-knowledge contexts. A common thread across these advances is that each utterance is represented by a single, temporally aggregated feature per modality. Our work departs from this convention by explicitly localizing intra-utterance emotion hotspots before performing fusion and alignment.

2.2 Multimodal Fusion

Multimodal sentiment and emotion analysis has explored a wide range of fusion strategies. Tensor Fusion (Zadeh et al., 2017) and its low-rank variant LMF (Liu et al., 2018) capture inter-modal interactions through outer products. Attention-based approaches such as MulT (Tsai et al., 2019)

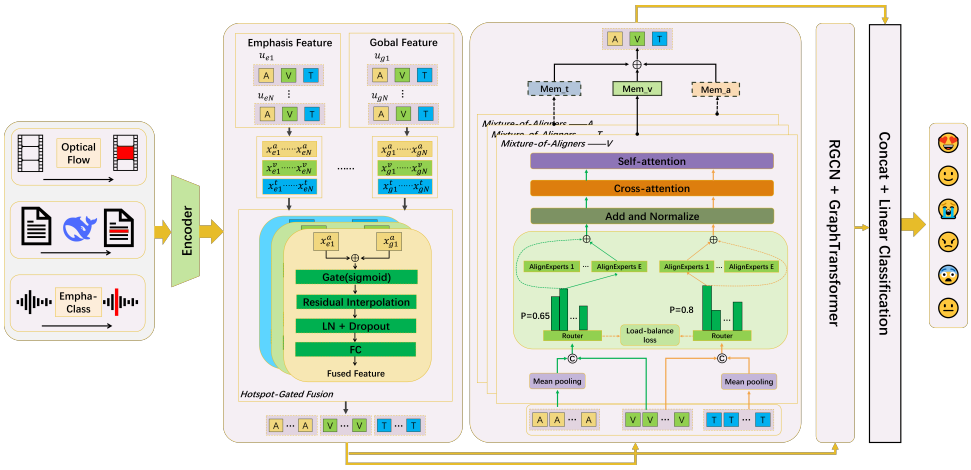

Figure 2: Per-utterance inputs from text (T), audio (A), and video (V) first undergo modality-specific **hotspot extraction** (LLM-based span selection, prosodic burst detection, and optical-flow analysis, respectively), producing hotspot sequences that are fused with global features by **HGF** and encoded into a shared representation. Two pathways operate in parallel: (i) **MoA** performs routed multi-expert cross-modal alignment and builds a modality memory; (ii) a **graph** models relational structure. The pathway outputs are concatenated and fed to a classifier. Implementation details are deferred to Section 3.

perform pairwise directional cross-modal attention, while MISA (Hazarika et al., 2020) decomposes representations into modality-invariant and modality-specific subspaces. MMIM (Han et al., 2021) adopts a mutual-information maximization objective across hierarchical fusion stages, and UniMSE (Hu et al., 2022) unifies sentiment and emotion tasks under a shared encoder–decoder. In conversational settings, GBAN (Liu et al., 2020) gates aligned and global features, and MMoE (Yu et al., 2024) routes inputs through specialized multimodal interaction experts. Despite their effectiveness, these methods operate on global utterance-level representations and do not explicitly distinguish high-salience emotional segments from neutral filler. Our HGF addresses this gap by performing token-wise gated interpolation between extracted hotspot features and global context.

2.3 Cross-Modal Alignment

Cross-modal alignment becomes particularly challenging when models operate at fine granularities. Because emotion hotspots across modalities rarely peak simultaneously—prosody may crest before the face reacts, and both may lead or lag the decisive word (Busso and Narayanan, 2007; Metallinou and Narayanan, 2013)—rigid temporal alignment forces the most informative segments to be paired with neutral or irrelevant content. Prior methods such as MulT (Tsai et al., 2019), MCTN (Pham et al., 2019), MAG (Rahman et al., 2020) apply uniform alignment over entire sequences, which is adequate for coarse global features but dilutes the sparse hotspot evidence that a granularity-aware framework elevates. Our MoA addresses this by following the sparse gating paradigm of MoE (Shazeer et al., 2017; Fedus et al., 2022), routing each utterance through a dynamically selected subset of lightweight expert cross-attention heads so that the alignment strategy adapts to each sample's cross-modal dynamics.

2.4 Salient Region Detection in Affective Computing

Localizing emotionally salient segments within individual modalities has been studied across domains. In vision, the Facial Action Coding System (Ekman and Friesen, 1978) established the link between transient muscle activations and expressed affect, motivating apex-frame and peak-frame spotting for facial expression analysis (Zhalehpour et al., 2014). In speech, local-attention mechanisms that focus on prosodically prominent regions outperform utterance-level pooling for emotion recognition (Mirsamadi et al., 2017), and self-supervised representations such as wav2vec 2.0 (Baevski et al., 2020) further improve acoustic feature quality. In text, emotion-cause span extraction (Poria et al., 2021) identifies the textual triggers of emotion within dialogues. However, these lines of work remain modality-specific and have not been integrated into a unified multimodal framework. Our

approach bridges this gap by combining modality-tailored hotspot extraction with a shared gated fusion and routed alignment pipeline.

3 Method

3.1 Overview

The complete architecture is illustrated in Figure 2. Given text, audio, and video for each utterance (with conversational context), our framework models emotion at progressively broader granularities through three stages. **Stage 1: Hotspot Extraction** (Section 3.2)—modality-tailored strategies localize the fine-grained emotional evidence within each modality, producing a hotspot sequence alongside the global content sequence. **Stage 2: Local–Global Fusion** (Section 3.3)—Hotspot-Gated Fusion adaptively interpolates between hotspot and global features at each token position, yielding multi-granularity unimodal representations $\{\mathbf{X}_m\}$. **Stage 3: Cross-Modal Re-alignment** (Section 3.4)—a routed Mixture-of-Aligners and a relation-aware graph pathway operate in parallel on the enriched representations, capturing cross-modal semantic correspondences and dialogue-level structural dependencies respectively; the two streams are concatenated and fed to a classifier for utterance-level prediction.

3.2 Hotspot Extraction

As the first stage of the granularity-aware pipeline, we localize fine-grained emotional evidence within each modality. For each utterance u and modality $m \in \{\mathrm{T}, \mathrm{A}, \mathrm{V}\}$, we prepare a global content sequence $\mathbf{C}_m$ and extract a hotspot sequence $\mathbf{H}_m$ that captures localized emotional salience. The shared principle across all three modalities is to surface the sparse, high-intensity segments that coarse temporal aggregation would otherwise average away.

Text. We leverage a large language model (Liu et al., 2024) to identify *salient spans*—affect-laden vocabulary, ironic markers, or sentiment-bearing phrases—from the current utterance and its conversational context. The LLM serves solely as a context-aware span selector and performs no emotion classification. **Audio.** We detect *prosodic bursts*—moments of abrupt change in energy, pitch, and duration—using EmphaClass (de Seyssel et al., 2023), localizing the intervals where vocal emotional expression peaks. **Video.** We identify motion-salient regions via optical-flow analysis (Beauchemin and Barron, 1995; Huang et al., 2021), capturing abrupt facial actions and gestural shifts that concentrate emotional information within otherwise static visual streams.

3.3 Hotspot-Gated Fusion

The choice of gated interpolation follows directly from the multi-granularity principle that motivates our framework. Hotspots are *sparse, localized refinements* of the global context—they concentrate the decisive emotional evidence that utterance-level aggregation dilutes, but they are not a separate, self-sufficient information source. The fusion mechanism must therefore encode this asymmetric, hierarchical relationship between granularity levels.

Given the global sequence $\mathbf{C}_m$ and the extracted hotspot sequence $\mathbf{H}_m$ for each modality m, HGF performs token-wise gated interpolation:

$$\begin{aligned} \alpha_m &= \sigma\big(W_m[\mathbf{C}_m \,\|\, \mathbf{H}_m] + b_m\big), \quad \alpha_m \in \mathbb{R}^{L\times 1}, \\ \mathbf{Z}_m &= \mathbf{C}_m + \alpha_m \odot (\mathbf{H}_m - \mathbf{C}_m), \end{aligned} \tag{1}$$

where L is sequence length and $\odot$ denotes broadcasted element-wise multiplication. A per-modality encoder Enc_m then maps $\mathbf{Z}_m$ to a shared hidden space: $\mathbf{X}_m = \mathrm{Enc}_m(\mathbf{Z}_m)$.

Gated interpolation achieves this by letting each token position smoothly shift from its coarse global representation toward the fine-grained hotspot evidence only where the learned gate α_m deems the local cue sufficiently salient, while preserving global context elsewhere. This design is structurally preferable to symmetric alternatives such as concatenation, which leaves the burden of discovering the hierarchical relationship entirely to downstream layers, or attention-based fusion, which learns position-wise weights but imposes no constraint on how the two granularity levels interact. By interpolating along a single hotspot-to-global axis, gated fusion encodes the inductive bias that fine-grained cues should refine—rather than override—the global contextual scaffold.

3.4 Cross-Modal Re-alignment

With multi-granularity unimodal representations $\{\mathbf{X}_m\}$ produced by HGF, the final stage integrates information across modalities. Because emotion hotspots across modalities rarely peak at the same moment, naïve temporal alignment would pair the fine-grained cues elevated by Stages 1–2 with irrelevant content from other modalities, squandering the multi-granularity enrichment. Two complementary pathways therefore operate in parallel: a routed

Mixture-of-Aligners captures cross-modal semantic correspondences while adapting to each utterance's temporal dynamics, and a relation-aware graph pathway encodes dialogue-level structural dependencies.

Mixture-of-Aligners. For each ordered modality pair $(j \leftarrow k)$, we instantiate E lightweight aligner experts:

$$f^{(e)}_{j\leftarrow k} : (\mathbf{X}_j[t], \mathbf{X}_k) \mapsto \mathbb{R}^h, \quad e = 1, \ldots, E. \quad (2)$$

We perform utterance-level routing: for each target utterance t, concatenate the target embedding with a mean-pooled source context and produce expert logits,

$$\begin{aligned} \boldsymbol{\ell}_t &= \mathrm{Router}\big([\mathbf{X}_j[t]; \mathrm{MeanPool}(\mathbf{X}_k)]\big) \in \mathbb{R}^E, \\ \boldsymbol{\pi}_t &= \mathrm{Softmax}\big(\mathrm{TopKMask}(\boldsymbol{\ell}_t, K)\big). \end{aligned} \quad (3)$$

TopK is implemented as a masked softmax; experts are densely evaluated and linearly mixed (gated sparsification). The expert mixture is fed into a shallow cross-attention block to restore capacity, yielding

$$\mathbf{H}_{j\leftarrow k} = \mathrm{CA}\left(\sum_{e=1}^{E} \pi_{t,e} f^{(e)}_{j\leftarrow k}(\mathbf{X}_j, \mathbf{X}_k), \mathbf{X}_k\right), \quad (4)$$

where $\mathrm{CA}(\cdot)$ denotes the cross-attention mechanism. For each target modality j, we concatenate aligned representations from all sources and apply self-attention to build a modality memory, then concatenate across all targets to form the cross-modal representation:

$$\begin{aligned} \mathbf{H}^{\mathrm{mem}}_j &= \mathrm{SelfAttn}\Big(\mathrm{concat}_{k\neq j}\,[\mathbf{H}_{j\leftarrow k}]\Big), \\ \mathbf{H}^{\mathrm{moa}} &= \mathrm{concat}_j\,[\mathbf{H}^{\mathrm{mem}}_j]. \end{aligned} \quad (5)$$

Graph Pathway. We build an utterance-level multi-relational graph whose nodes are (modality m, utterance t). Edges consist of intra-modal temporal links within a window and optional same-time cross-modal links; each edge is typed by its temporal direction $s \in \{-1, 0, 1\}$ and the source/target modality pair. A relation-aware GNN—integrating an RGCN (Schlichtkrull et al., 2018) to model heterogeneous interactions and a Graph Transformer (Yun et al., 2020) to capture complex patterns—produces node embeddings that are concatenated per utterance across modalities:

$$\mathbf{H}^{\mathrm{gnn}} = \mathrm{MC}\big(\mathrm{GNN}(\mathbf{X}^{\mathrm{graph}}, \mathcal{E}),\ \mathrm{lengths},\ |\mathcal{M}|\big), \quad (6)$$

where $\mathcal{E}$ is the set of typed edges, and $\mathrm{MC}(\cdot)$ denotes the Multi-Concat operation that restores the graph node embeddings into utterance-wise representations across $|\mathcal{M}|$ modalities.

3.5 Loss Function

We optimize a sum of the utterance-level task loss and an optional MoA load-balancing regularizer:

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{task}} \;+\; \lambda\,\mathcal{L}_{\mathrm{lb}}. \quad (7)$$

Here $\lambda \geq 0$ is a scalar weight (we use $\lambda > 0$ by default).

The task loss is the negative log-likelihood over utterances:

$$\mathcal{L}_{\mathrm{task}} \;=\; -\,\mathbb{E}_{(b,t)}\big[\log p(y_{b,t} \mid \mathbf{H}_{b,t})\big], \quad (8)$$

where (b, t) indexes batch b and utterance t, $\mathbf{H}_{b,t}$ is the model's utterance representation (concatenating the MoA and graph pathways), and $p(\cdot \mid \mathbf{H}_{b,t})$ is produced by the classifier; class-balancing weights can be applied when needed.

The load-balancing term encourages uniform expert usage in MoA:

$$\begin{aligned} \mathcal{L}_{\mathrm{lb}} \;&=\; \sum_{e=1}^{E} u_e \log u_e \;+\; \log E, \\ u_e \;&=\; \mathbb{E}_{(b,t)}\big[\pi_{t,e}\big]. \end{aligned} \quad (9)$$

Here E is the number of experts, $\pi_{t,e}$ is the Top$K$-masked softmax routing weight, and u_e is the average expert usage.

4 Experiments

4.1 Experimental Setting

Datasets. We evaluate our method on two public multimodal ERC datasets: IEMOCAP (Busso et al., 2008) and CMU-MOSEI (Bagher Zadeh et al., 2018). IEMOCAP contains 151 dialogues (7433 utterances, 12 hours in total) annotated with six emotions: happy, sad, neutral, angry, excited, and frustrated. To reduce label ambiguity, (happy, excited) and (sad, frustrated) are merged into a four-class setting. CMU-MOSEI provides annotations for six basic emotions: happiness, sadness, anger, fear, surprise, and disgust.

Evaluation Metrics. The primary metrics for evaluation are the *weighted F1-score* (w-F1) and *Accuracy* (Acc.). The w-F1 metric is computed as the sum $\sum_{k=1}^{K} \mathrm{freq}_k \times \mathrm{F1}_k$, where freq_k is the relative frequency associated with class k. Accuracy,

Table 1: Comparison with previous state-of-the-art methods on IEMOCAP (6-way). The **bolded** results indicate the best performance, while the underlined results represent the second best performance.

Methods	IEMOCAP(6-way)							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc.(%)	w-F1(%)
DialogueRNN (Majumder et al., 2019)	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75
DialogueGCN (Ghosal et al., 2019)	47.10	80.88	58.71	66.08	70.97	61.21	65.54	65.04
DialogueCRN (Hu et al., 2021)	51.59	74.54	62.38	67.25	73.96	59.97	65.31	65.34
MMGCN (Wei et al., 2019)	45.45	77.53	61.99	66.70	72.04	64.12	65.56	65.71
COGMEN (Joshi et al., 2022)	55.76	80.17	63.21	61.69	74.91	63.90	67.04	67.27
CORECT (Nguyen et al., 2023)	56.19	82.11	63.97	66.49	69.12	61.91	66.61	66.74
M3NET (Chen et al., 2023)	56.77	82.26	68.94	66.67	78.06	60.32	69.38	69.29
CMERC (Tu et al., 2024)	**60.73**	81.89	71.65	**69.51**	77.45	**67.02**	-	71.98
Ours	58.97	**83.67**	**72.35**	69.23	**79.13**	66.94	**72.52**	**72.52**

Table 2: Results on IEMOCAP (4-way) and CMU-MOSEI dataset compared with previous works. For sentiment classification on the CMU-MOSEI dataset, we trained an independent binary classifier for each emotion. The **bolded** results indicate the best performance, while the underlined results represent the second best performance. * denotes identical scores for Fear (84.47) and Surprise (85.66) due to the baselines' binary classifiers failing to separate the positive and negative classes.

Methods	IEMOCAP(4-way)		CMU-MOSEI (w-F1(%))					
	Acc.(%)	w-F1(%)	Happiness	Sadness	Angry	Fear	Disgust	Surprise
bc-LSTM (Poria et al., 2017)	75.20	75.13	-	-	-	-	-	-
CHFusion (Majumder et al., 2018)	76.59	76.80	-	-	-	-	-	-
Multilouge-Net (Shenoy and Sardana, 2020)	-	-	67.84	65.34	67.03	84.47*	74.91	85.66*
TBJE (Delbrouck et al., 2020)	-	-	65.91	68.54	70.78	84.47*	80.23	85.66*
COGMEN (Joshi et al., 2022)	82.29	82.15	68.73	69.11	74.20	84.47*	80.23	85.66*
CORECT (Nguyen et al., 2023)	82.93	82.92	69.24	70.67	76.36	84.47*	82.84	**86.00**
Ours	**85.90**	**85.89**	**71.04**	**72.34**	**76.43**	**84.81**	**85.21**	85.84

conversely, measures the percentage of correct predictions over the entire test set.

Baselines and Implementation. Our model extends the CORECT framework (Nguyen et al., 2023) with modifications while ensuring fair comparison with baseline methods. On IEMOCAP, we conduct both six-class and four-class experiments, while on CMU-MOSEI, we perform binary classification for each emotion category. The implementation is based on PyTorch 2.1.1 and torch-geometric, trained on 4×L20 GPUs. We use the Adam optimizer with a learning rate of 9×10^{-5} and a ReduceLR scheduler. The hidden dimension is set to 100 with a dropout rate of 0.48. For the proposed MoA module, we use 6 experts with top-3 gating. Full hyperparameter settings are provided in Appendix A.

4.2 Overall Results

We compare our model against dataset-specific SOTA baselines spanning RNN-based, graph-based, and multimodal fusion methods. Tables 1 and 2 report the full results.

IEMOCAP (6-way). Our model achieves 72.52/72.52 (Acc./w-F1), setting a new state of the art on both metrics. The largest per-class gains over CORECT appear on *Neutral* (+8.38) and *Excited* (+10.01)—two categories most prone to inter-class confusion. These gains suggest that explicit hotspot extraction surfaces the fine-grained temporal cues—transient prosodic bursts distinguishing excitement from happiness, fleeting facial details separating composure from suppressed frustration—that global pooling averages away, while routed cross-modal alignment ensures such cues reinforce each other across modalities.

IEMOCAP (4-way). Merging (Happy, Excited) and (Sad, Frustrated) into a four-class setting, our model reaches 85.90/85.89 (Acc./w-F1), outperforming CORECT by +2.97. The smaller absolute gain relative to the six-class setting is expected, as four-class labeling collapses the very pairs whose boundaries benefit most from hotspot-level discrimination.

CMU-MOSEI. Our model leads on five of six emotion categories (Table 2), with the strongest gains on *Disgust*, *Happiness*, and *Sadness* over CORECT—emotions whose discriminative cues of-

Table 3: Ablation study on IEMOCAP (6-way). The **bolded** results indicate the best performance, while the underlined results represent the second best performance.

Methods	IEMOCAP(6-way)							
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc.(%)	w-F1(%)
baseline	56.88	79.19	64.34	64.62	70.59	63.23	66.67	66.84
baseline (Transformer-based)	52.60	76.99	67.58	65.23	71.21	62.70	66.61	66.95
baseline + HGF (without text)	**59.56**	79.58	69.74	63.64	71.91	62.53	68.21	68.39
baseline + HGF	56.69	80.33	70.48	68.91	78.25	63.45	70.67	70.36
Full Models	58.97	**83.67**	**72.35**	69.23	**79.13**	66.94	**72.52**	**72.52**

ten reside in brief, low-intensity prosodic or facial signals that temporal aggregation dilutes, making them particularly amenable to explicit hotspot extraction. On *Fear*, our model is the first to surpass the ceiling shared by all prior baselines, suggesting that adaptive cross-modal routing captures asynchronous vocal–visual fear cues that rigid alignment discards. *Surprise* is the only category where we rank second, marginally below CORECT.

4.3 Ablation Study

As Table 3 shows, we ablate modules incrementally atop the original baseline, which already includes the cross-modal graph pathway. We keep the graph pathway intact for fairness to baseline (Nguyen et al., 2023) and because our design complements—rather than replaces—its relational/temporal modeling.

Adding HGF provides the first substantial improvement by amplifying emotion hotspot-guided local cues and reducing reliance on aggressive pooling, with table evidence of stronger Neutral/Excited performance and overall stability. Adding MoA on top delivers further, steady gains by selectively routing cross-modal evidence to the target utterance and sharpening alignment where it is most discriminative, reflected in the consistent transition from baseline to +HGF to +HGF+MoA in both aggregate and class-wise results.

Implicit attention is not enough. Replacing the baseline's fully-connected fusion layers with Transformer blocks (*baseline (Transformer-based)* in Table 3) yields no meaningful improvement, despite the attention mechanism's theoretical ability to up-weight salient positions. We attribute this to the sparsity of emotion hotspots: the affectively decisive segments occupy only a small fraction of each utterance and are easily overwhelmed by neutral tokens and surface-similar distractors. Without an explicit prior indicating *where* to look, implicit attention cannot reliably isolate sparse emotional evidence, motivating our explicit hotspot extraction and gated fusion design.

Cross-modal contribution analysis. To examine whether the performance gains are driven by all modalities rather than text alone, we ablate text-modality hotspots entirely (*baseline + HGF (without text)* in Table 3), retaining only audio and video hotspot extraction. The resulting model reaches 68.21/68.39 (Acc./w-F1), already outperforming the full three-modality baseline (66.67/66.84) by a clear margin. This confirms that audio and video hotspots independently provide substantial discriminative gains through the multi-granularity pipeline. Reintroducing text hotspots further improves performance to 70.67/70.36, demonstrating that each modality contributes complementary fine-grained evidence. We note that the LLM employed for text span selection serves solely as a context-aware segment locator—performing no emotion classification and introducing no training data beyond the original dialogues—consistent with the observation that the framework delivers strong gains even without any text-side hotspot extraction.

Confusion-matrix analysis. Figure 3 compares the confusion patterns of the baseline and the full model on IEMOCAP (6-way). The largest diagonal gains appear on *Excited* (+8.3 pp) and *Neutral* (+8.1 pp), consistent with the per-class F1 trends above. These gains are driven by reduced confusion in the two most error-prone off-diagonal pairs: the Excited→Happy misclassification rate drops by 4.4 pp, and Neutral→Frustrated by 3.4 pp. We attribute this improved separability to the multi-granularity design: HGF surfaces fine-grained hotspot cues—such as transient prosodic bursts that distinguish genuine excitement from happiness, or fleeting facial micro-expressions that separate neutral composure from suppressed frustration—that global pooling averages away; MoA then selectively routes cross-modal evidence through specialized alignment experts, ensuring that corroborative

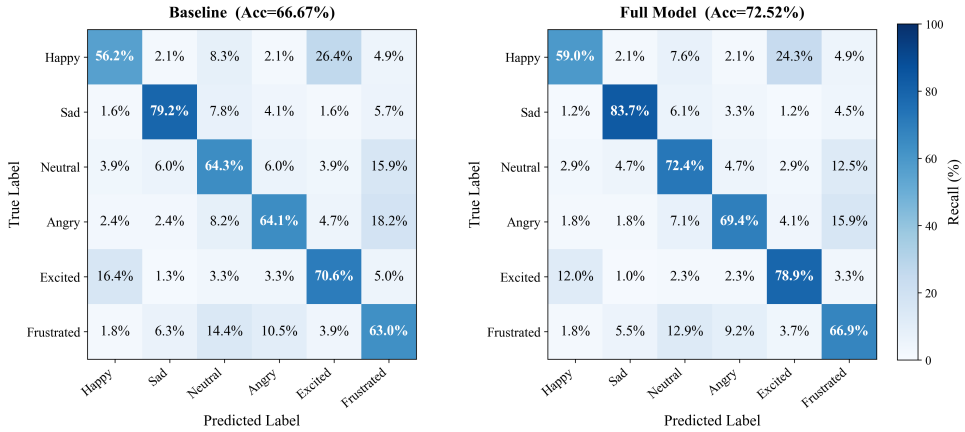

Figure 3: Confusion matrix of baseline and the proposed model on IEMOCAP (6-way).

cues across modalities reinforce rather than dilute each other. The broad decrease in off-diagonal values confirms that modeling emotion at multiple granularities yields more discriminative representations, with the strongest error reductions on the most confusable pairs—also explaining why the gain narrows in the four-class setting (Table 2), where these pairs are merged.

Table 4: Effect of re-alignment diversity on IEMOCAP (4-way): expert pool size E controls the repertoire of available alignment strategies in Stage 3, while top-K governs how many are composed per utterance. The **bolded** result indicates the best performance.

Configuration (E, K)	w-F1 (%)
$E = 4, K = 2$	80.92
$E = 4, K = 3$	81.25
$E = 6, K = 2$	82.31
$\mathbf{E = 6, K = 3}$	**85.89**

Re-alignment diversity and selectivity. Table 4 examines how the expert pool size E and the number of activated experts K in MoA affect Stage 3 performance on IEMOCAP (4-way). A larger expert pool consistently improves results by offering the router a richer repertoire of alignment strategies, each specializing in different cross-modal interaction patterns over the hotspot-enriched representations from HGF. More notably, the benefit of raising K depends strongly on E: this interaction is governed by the effective sparsity ratio K/E—when the ratio is high, nearly all experts are activated, eliminating the router's ability to selectively compose the most relevant aligners for each utterance; a lower ratio preserves meaningful sparsity, letting the router tailor re-alignment to each sample's specific cross-modal dynamics. These results confirm that the re-alignment stage benefits from both diversity (more experts) and selectivity (sparse routing), paralleling the multi-granularity principle that drives the overall framework: just as HGF balances fine-grained and global cues within each modality, MoA routing balances breadth and specificity across modalities.

5 Conclusion

We identify a fundamental *granularity gap* in existing ERC methods: utterance-level temporal aggregation discards the fine-grained emotional evidence that is often decisive for discriminating confusable emotions. To bridge this gap, we propose a granularity-aware framework organized in three stages. First, modality-tailored hotspot extraction localizes the sparse, high-intensity segments where emotions peak. Second, Hotspot-Gated Fusion adaptively interpolates between these fine-grained hotspots and global utterance context, yielding multi-granularity unimodal representations that preserve both precise local cues and the broader scaffold for disambiguation. Third, a routed Mixture-of-Aligners flexibly re-aligns the enriched cross-modal representations, complemented by a graph pathway that encodes conversational structure. Experiments on IEMOCAP and CMU-MOSEI demonstrate state-of-the-art performance, and ablations confirm that each granularity stage contributes complementary gains—with explicit hotspot extraction proving essential where implicit attention fails. These results suggest that modeling emotion at the granularity at which it actually occurs, rather than collapsing all temporal positions into a single vector, offers a principled and effective paradigm for multimodal ERC.

Limitations

Hotspot extraction is performed offline and no ground-truth hotspot annotations exist for the benchmarks we use, so the heuristic extraction strategies are validated only indirectly through end-task performance. Our evaluation is further constrained by English-language datasets with limited ecological validity, leaving both cross-lingual and in-the-wild generalization untested. Additionally, the model assumes the availability of all three modalities; graceful degradation under missing-modality conditions is not explored.

Ethics Statement

This work uses two publicly available datasets, IEMOCAP (Busso et al., 2008) and CMU-MOSEI (Bagher Zadeh et al., 2018), both collected with informed consent under their respective institutional protocols; no additional human data was collected or annotated. A large language model is employed solely as a span selector for text hotspot extraction and performs no emotion classification, so its use introduces no external labeling bias. We acknowledge that emotion recognition technology carries potential risks of misuse in surveillance or privacy-invasive applications; we encourage its deployment only in consensual, user-beneficial settings.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, and 1 others. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the ACL*, pages 2236–2246.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68.

Steven S. Beauchemin and John L. Barron. 1995. The computation of optical flow. *ACM computing surveys (CSUR)*, pages 433–466.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, and 1 others. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, pages 335–359.

Carlos Busso and Shrikanth S Narayanan. 2007. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2331–2347.

Feiyu Chen, Jie Shao, Shuyuan Zhu, and 1 others. 2023. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *2023 IEEE/CVF CVPR*, pages 10761–10770.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and 1 others. 2023. Emphassess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. *arXiv preprint arXiv:2312.14069*.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and 1 others. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7.

Paul Ekman. 2003. Darwin, deception, and facial expression. *Annals of the new York Academy of sciences*, 1000(1):205–221.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, pages 1–39.

Yao Fu, Shaoyang Yuan, Chi Zhang, and 1 others. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, and 1 others. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the EMNLP-IJCNLP*, pages 154–164.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 9180–9192.

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

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the ACL*, pages 7042–7052, Online.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 7837–7851.

Shih-Chia Huang, Trung-Hieu Le, and Da-Wei Jaw. 2021. Dsnet: Joint semantic learning for object detection in inclement weather conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2623–2633.

Peiyuan Jiang, Yao Liu, Qiao Liu, Zongshun Zhang, Jiaye Yang, Lu Liu, and Daibing Yao. 2025. Drkf: Decoupled representations with knowledge fusion for multimodal emotion recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 955–964, New York, NY, USA. Association for Computing Machinery.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, and 1 others. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the ACL*, pages 4148–4164.

Mustaqeem Khan, Phuong-Nam Tran, Nhat Truong Pham, Abdulmotaleb El Saddik, and Alice Othmani. 2025. Memocmt: multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific Reports*, 15(1):5473.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*.

Xin Li, Xin He, Gaigai Tang, Kaiyuan Qi, Guangfeng Su, and Huiyun Zhang. 2026. Discrepancy-aware contrastive learning with mixture of experts for cross-modal image-text semantic alignment. *Neurocomputing*, 660:131887.

Aixin Liu, Bei Feng, Bing Xue, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Pengfei Liu, Kun Li, and Helen Meng. 2020. Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition. In *INTERSPEECH*, pages 379–383.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

Samaneh Madanian, Talen Chen, Olayinka Adeleye, and 1 others. 2023. Speech emotion recognition using machine learning — a systematic review. *Intelligent Systems with Applications*, page 200266.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, and 1 others. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, 161:124–133.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, and 1 others. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6818–6825.

Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2227–2231. IEEE.

Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, and 1 others. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the EMNLP*, pages 15154–15167.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6892–6899.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and 1 others. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the ACL*, pages 873–883.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, and 1 others. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2359–2369.

Klaus R Scherer. 2009. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3459–3474.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, and 1 others. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the ACL*.

Geng Tu, Feng Xiong, Bin Liang, and 1 others. 2024. Multimodal emotion recognition calibration in conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 9621–9630.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

Yan Wang, Wei Song, Wei Tao, and 1 others. 2022. A systematic review on affective computing: emotion models, databases, and recent advances. *Information Fusion*, pages 19–52.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Yinwei Wei, Xiang Wang, Liqiang Nie, and 1 others. 2019. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, page 1437–1445.

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

Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. 2013. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of nonverbal behavior*, 37(4):217–230.

Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, and 1 others. 2024. MMoE: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the EMNLP*, pages 10006–10030.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. 2020. Graph transformer networks. *Preprint*, arXiv:1911.06455.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1103–1114.

Sara Zhalehpour, Zahid Akhtar, and Cigdem Eroglu Erdem. 2014. Multimodal emotion recognition with automatic peak frame selection. In *2014 IEEE International Symposium on Innovations in INISTA Proceedings*, pages 116–121. IEEE.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2025. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, page 107901.

A Hyperparameter Settings

Hyperparameter	Value
Hidden dimension	100
Batch size	10
Epochs	40
Learning rate	9×10^{-5}
Weight decay	5×10^{-8}
Dropout rate	0.48
Gradient clipping	1.0
Past / Future window	11 / 9
Encoder layers	2
Cross-modal attention	2 layers, 2 heads
Self-attention	3 layers, 2 heads
Graph Transformer heads	7
MoA experts	6
MoA top-k	3
Random seed	3407

Table 5: Hyperparameter settings for the proposed model.