HEAR: HIERARCHICAL EVIDENCE-AWARE REASONING VIA TEXT CONDITIONING FOR MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sentiment analysis (MSA) seeks to infer human sentiment from textual, acoustic, and visual signals. Existing approaches often rely on feature fusion or cross-modal interaction, where globally salient information may overshadow subtle yet sentiment-discriminative cues. This highlights a key challenge: perceptual saliency does not necessarily imply sentiment relevance, particularly when modalities convey discrepant information. To address this challenge, we propose **HEAR**, a **H**ierarchical **E**vidence-**A**ware **R**easoning framework that shifts multimodal sentiment modeling from direct feature fusion toward structured evidence reasoning. HEAR first extracts fine-grained candidate evidence from complementary temporal and spectral views of acoustic and visual features while preserving shared and view-specific information. It then conditions the evidence on textual semantics to distinguish text-aligned from text-deviating evidence, explicitly preserving both cross-modal consistency and discrepancy. Rather than directly aggregating evidence into a single global representation, HEAR progressively reasons from local cues through relational interactions to global sentiment abstractions. Hierarchical evidence is incrementally projected into pseudo-tokens and integrated with textual embeddings for sentiment prediction using a frozen large language model. Experiments on CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA demonstrate consistent improvements over strong fusion-based and LLM-based baselines.

1 INTRODUCTION

Multimodal sentiment analysis (MSA) aims to understand human affect by jointly exploiting textual, acoustic, and visual signals (Gandhi et al., 2023; Das & Singh, 2023; Fang et al., 2025). Compared with text-only sentiment analysis (Kang et al., 2023; Li et al., 2021), multimodal observations provide complementary cues, motivating research on feature-level fusion (Cheng et al., 2023; Liu et al., 2018) and cross-modal interaction (Zhao et al., 2026; Zhuang et al., 2024; Ou & Li, 2026). More recently, LLM-based methods (Yang et al., 2025a; Liu et al., 2025; Li et al., 2024c) have incorporated non-textual information into pretrained language models. Despite these advances, most methods remain representation-centric, focusing on learning unified representations through cross-modal alignment, interaction, or aggregation.

Motivation. However, *perceptual saliency does not necessarily imply sentiment relevance*. As shown in Fig. 1, salient cues such as lively intonation or a smile may dominate multimodal representations yet mislead sentiment prediction, while subtle but discriminative cues may be obscured by global aggregation. Moreover, sentiment relevance is not equivalent to cross-modal agreement: agreement with text may reinforce sentiment, whereas deviations can provide complementary evidence. Treating such discrepancies as noise may therefore discard useful sentiment cues. These observations motivate a shift from **multimodal representation fusion** to **multimodal evidence reasoning**, raising three challenges: (i) discovering fine-grained acoustic and visual evidence while preserving subtle cues; (ii) contextualizing such evidence with linguistic semantics while retaining both agreement and informative deviation; and (iii) progressively organizing heterogeneous evidence from individual cues to their interactions and global abstractions.

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.

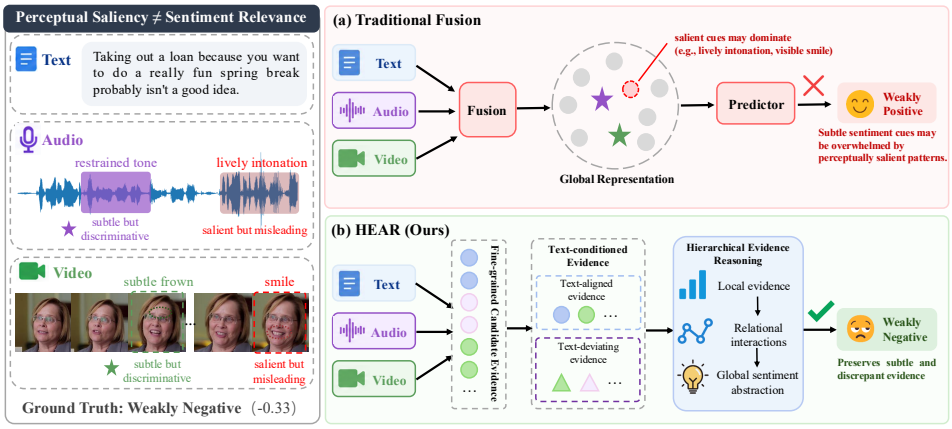


Figure 1: **Motivation and comparison.** (a) Traditional fusion may overlook subtle sentiment cues in favor of perceptually salient but misleading signals. (b) HEAR preserves text-aligned and text-deviating evidence through hierarchical reasoning, enabling informed sentiment prediction.

Our Approach. To address these challenges, we propose **HEAR**, a **H**ierarchical **E**vidence-**A**ware **R**easoning framework via text conditioning for MSA. As illustrated in Fig. 1(b), HEAR progressively discovers, interprets, and reasons over sentiment evidence rather than directly constructing a unified multimodal representation. **Fine-grained evidence discovery.** HEAR extracts candidate evidence from complementary temporal and spectral views while preserving shared and view-specific information. **Text-conditioned evidence interpretation.** HEAR uses textual semantics to construct both text-aligned and text-deviating evidence, preserving potentially informative discrepancies. **Hierarchical evidence reasoning.** HEAR progressively organizes evidence from local cues to relational interactions and global abstractions, which are incrementally projected into pseudo-tokens for a frozen LLM. Experiments on four benchmarks demonstrate consistent improvements over strong fusion-based and LLM-based baselines. Our main contributions are summarized as follows:

- **Evidence Reasoning Perspective.** We revisit MSA from an evidence reasoning perspective, shifting the focus from constructing unified multimodal representations toward discovering, interpreting, and reasoning over fine-grained sentiment evidence.
- **Evidence Discovery and Interpretation.** We propose **HEAR**, which discovers candidate evidence from complementary temporal–spectral views and interprets it through text conditioning, explicitly preserving both text-aligned and text-deviating evidence.
- **Hierarchical Evidence Reasoning.** We progressively organize evidence from local cues to relational interactions and global abstractions, together with an incremental pseudo-token projection for integrating structured evidence into a frozen LLM.
- **Empirical Evaluation.** Extensive experiments on four multimodal affective benchmarks demonstrate consistent improvements over strong baselines, with targeted analyses validating HEAR under subtle sentiment cues and cross-modal discrepancies.

2 Related Work

Multimodal Sentiment Analysis. Existing MSA studies have primarily advanced through multimodal fusion and representation learning. Fusion-based methods integrate heterogeneous modalities via tensor-based fusion (Zadeh et al., 2017) or cross-modal interaction (Tsai et al., 2019; Han et al., 2021; Li et al., 2024b) to capture multimodal dependencies. In parallel, representation-based approaches learn modality-invariant and modality-specific representations (Hazarika et al., 2020; Yu et al., 2021; Yang et al., 2022), disentangle shared and private factors (Yu et al., 2023b; Li et al., 2023b; 2026), or explore fine-grained cross-modal patterns (Zhuang et al., 2024; Li et al., 2026). Despite these advances, fusion and representation learning do not explicitly distinguish perceptual saliency from sentiment relevance, potentially obscuring subtle yet discriminative cues under cross-modal discrepancy. HEAR instead shifts the focus from multimodal representation learning and fusion toward structured reasoning over fine-grained sentiment evidence.

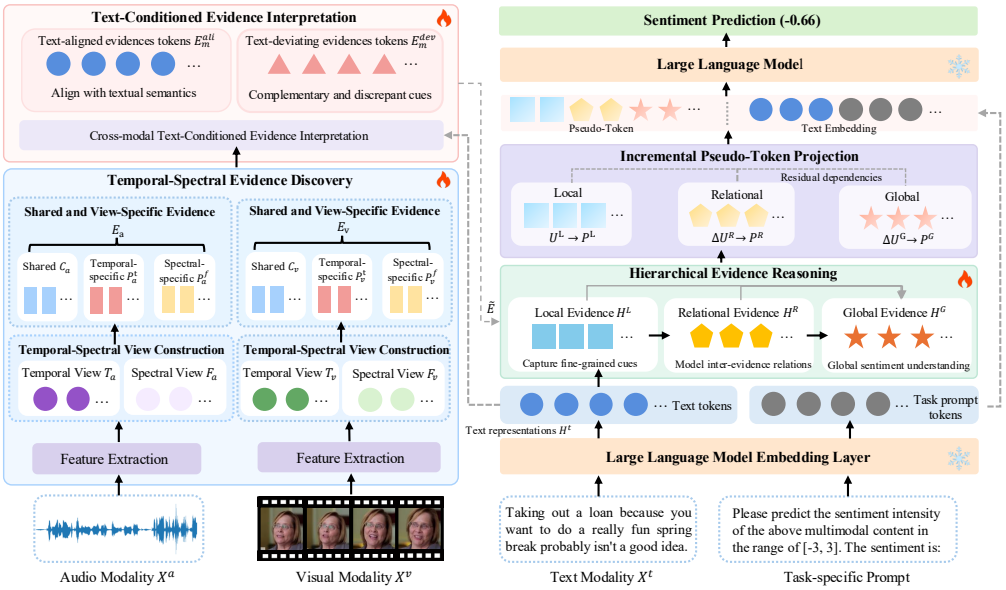


Figure 2: Overall architecture of the proposed HEAR framework. Temporal–spectral evidence discovery and text-conditioned interpretation construct complementary multimodal evidence. Hierarchical evidence reasoning and incremental pseudo-token projection progressively organize the evidence and inject it into a frozen LLM for sentiment prediction.

Multimodal Large Language Models. Pre-trained language models (PLMs) have been extended to multimodal sentiment understanding by integrating non-textual features (Guo et al., 2022), translating acoustic and visual signals into text (Wu et al., 2025), or adopting multimodal pre-training objectives (Hu et al., 2022b; Lin et al., 2025). With the emergence of LLMs, recent studies increasingly employ lightweight interfaces, including projection-based mappings (Eichenberg et al., 2022; Liang et al., 2022), learnable queries (Alayrac et al., 2022; Li et al., 2023a), adapters (Yang et al., 2025b; Chen et al., 2025), and soft prompts or pseudo-tokens (Liang et al., 2024; Liu et al., 2026), to bridge non-textual modalities with language backbones. While these methods primarily address how multimodal representations are delivered to LLMs, less attention is paid to what sentiment-relevant evidence should be preserved and how it should be organized beforehand. HEAR addresses this gap by structuring fine-grained multimodal cues into sentiment evidence prior to LLM reasoning.

3 METHODS

Problem Formulation and Overview. Given a multimodal utterance $\mathcal{X} = \{X^t, X^a, X^v\}$, where X^t denotes text and $X^m \in \mathbb{R}^{T_m \times d_m}$ denotes the feature sequence for $m \in \{a, v\}$, MSA predicts the sentiment label or intensity as $\hat{y} = f_\Theta(\mathcal{X})$. As illustrated in Fig. 2, HEAR comprises four stages. *Temporal–Spectral Evidence Discovery* extracts shared and view-specific evidence from acoustic and visual features. *Text-Conditioned Evidence Interpretation* constructs text-aligned and text-deviating evidence. *Hierarchical Evidence Reasoning* organizes the evidence from local cues to relational interactions and global abstractions. Finally, *Incremental Pseudo-Token Projection* converts hierarchical evidence into pseudo-tokens for integration with text and a frozen LLM.

3.1 TEMPORAL–SPECTRAL EVIDENCE DISCOVERY

To preserve fine-grained multimodal cues, we construct temporal and spectral views for each non-textual modality and decompose them into shared and view-specific representations. Rather than compressing these representations into a single global vector, we retain them as fine-grained candidate evidence for subsequent text-conditioned interpretation.

Temporal–Spectral View Construction. For each modality $m \in \{a, v\}$, let $X_m \in \mathbb{R}^{L_m \times d_m}$ denote its valid feature sequence. We resample X_m to K temporal positions, project it into a d-dimensional space, and add learnable positional embeddings before temporal encoding:

$$T_m = \mathrm{Transformer}_t(\mathrm{Proj}_t(\mathrm{Resample}_K(X_m)) + P_m)\,, \tag{1}$$

where $P_m \in \mathbb{R}^{K \times d}$ and $T_m \in \mathbb{R}^{K \times d}$ denote the positional embeddings and temporal representation, respectively. For the spectral view, we apply a Hann window and compute the log-power spectrum $S_m = \log(1 + |\,\mathrm{rFFT}_t(w_m \odot X_m)|^2)$. After interpolating it to N_f frequency bins and projecting it to dimension d, K learnable queries aggregate frequency information via cross-attention:

$$F_m = \mathrm{Transformer}_f\Big(\mathrm{CrossAttn}\Big(Q_m, \mathrm{Proj}_f(\mathrm{Interp}_{N_f}(S_m)), \mathrm{Proj}_f(\mathrm{Interp}_{N_f}(S_m))\Big)\Big)\,, \tag{2}$$

where $Q_m, F_m \in \mathbb{R}^{K \times d}$ denote the learnable queries and spectral representation, respectively. This branch captures frequency variation over the entire valid sequence rather than temporally localized time–frequency patterns.

Shared and View-Specific Evidence Modeling. To capture overlapping and complementary information between temporal and spectral views, we perform a shared–private decomposition within each modality. The two views are projected into a common latent space, where the shared component is obtained by slot-wise averaging and normalization, and view-specific components by subtracting it from independently projected views:

$$(C_m, P_m^t, P_m^f) = \left(\mathrm{LN}\left(\frac{C_m^t + C_m^f}{2}\right), \mathrm{LN}\big(W_m^{tp} T_m + b_m^{tp} - C_m\big), \mathrm{LN}\big(W_m^{fp} F_m + b_m^{fp} - C_m\big)\right), \tag{3}$$

where $C_m^t = W_m^{tc} T_m + b_m^{tc}$ and $C_m^f = W_m^{fc} F_m + b_m^{fc}$. Here, C_m, P_m^t, and P_m^f represent shared, temporal-specific, and spectral-specific information, respectively. Since temporal tokens and spectral queries lack inherent one-to-one physical correspondence, their slot-wise combination represents latent-slot alignment rather than explicit temporal–frequency alignment.

Candidate Evidence Construction. We retain the shared and view-specific components as separate candidate evidence and concatenate them along the token dimension:

$$E_m = \mathrm{Concat}_{\mathrm{tok}}\big(C_m, P_m^t, P_m^f\big) \in \mathbb{R}^{3K \times d}, \tag{4}$$

The discovery stage thus produces $\mathcal{E}_{\mathrm{disc}} = \{E_a, E_v\}$ with $6K$ candidate evidence tokens. Each token represents a latent candidate cue rather than an explicitly localized segment, semantic description, or predefined sentiment category; its sentiment relevance is subsequently determined through text-conditioned interpretation. To preserve the shared–private decomposition while reducing redundancy, we impose reconstruction, shared consistency, and decorrelation constraints:

$$\begin{cases} \mathcal{L}_{\mathrm{rec}}^m = \mathrm{MSE}\big(D_m^t([C_m; P_m^t]), T_m\big) + \mathrm{MSE}\big(D_m^f([C_m; P_m^f]), F_m\big)\,, \\ \mathcal{L}_{\mathrm{con}}^m = \mathrm{MSE}(C_m^t, C_m^f), \\ \mathcal{L}_{\mathrm{div}}^m = \dfrac{1}{3}\big[\mathrm{Decorr}(P_m^t, P_m^f) + \mathrm{Decorr}(C_m, P_m^t) + \mathrm{Decorr}(C_m, P_m^f)\big]\,, \end{cases} \tag{5}$$

where $\mathrm{Decorr}(U, V) = \frac{1}{K}\sum_{k=1}^{K} \cos^2(U_k, V_k)$ measures token-wise redundancy. The overall discovery loss is represented as:

$$\mathcal{L}_{\mathrm{disc}} = \frac{1}{2} \sum_{m \in \{a,v\}} (\mathcal{L}_{\mathrm{rec}}^m + \mathcal{L}_{\mathrm{con}}^m + \mathcal{L}_{\mathrm{div}}^m)\,, \tag{6}$$

with the three terms equally weighted and its overall contribution controlled by the auxiliary-loss weight in the training objective.

3.2 Text-Conditioned Evidence Interpretation

Candidate evidence from acoustic and visual signals does not inherently convey sentiment, and its interpretation depends on linguistic context. Moreover, deviations from textual semantics may provide complementary affective cues. We therefore use text as a semantic reference to construct both text-aligned and text-deviating evidence.

Textual Semantic Conditioning. Let $H^t \in \mathbb{R}^{L_t \times d_t}$ denote the textual representations extracted from the language model. Given candidate evidence $E_m \in \mathbb{R}^{3K \times d}$ for $m \in \{a, v\}$, we retrieve relevant textual context via cross-attention:

$$C_m^{\text{txt}} = \text{MHA}\left(E_m, W_t H^t, W_t H^t\right), \tag{7}$$

where W_t projects text representations into the d-dimensional evidence space, and $C_m^{\text{txt}} \in \mathbb{R}^{3K \times d}$ provides the corresponding textual context.

Text-Aligned Evidence Modeling. We combine each candidate evidence token with its retrieved textual context:

$$E_m^{\text{ali}} = \text{LN}\left(E_m + C_m^{\text{txt}}\right) + e_{\text{ali}}, \tag{8}$$

where $e_{\text{ali}} \in \mathbb{R}^d$ is a learnable branch embedding. This additive interaction reinforces text-contextualized information without explicitly discarding individual cues.

Text-Deviating Evidence Modeling. In parallel, we preserve deviations from the retrieved textual context through a residual interaction:

$$E_m^{\text{dev}} = \text{LN}\left(E_m - C_m^{\text{txt}}\right) + e_{\text{dev}}, \tag{9}$$

where $e_{\text{dev}} \in \mathbb{R}^d$ is the corresponding branch embedding. Such deviations reflect differences from textual semantics rather than necessarily indicating sentiment contradiction. We retain both branches and concatenate the interpreted acoustic and visual evidence:

$$\widetilde{E} = \text{Concat}_{\text{tok}}\left(E_a^{\text{ali}}, E_a^{\text{dev}}, E_v^{\text{ali}}, E_v^{\text{dev}}\right) \in \mathbb{R}^{12K \times d}. \tag{10}$$

Modality embeddings are added to distinguish acoustic and visual evidence. The resulting evidence bank preserves both text-aligned and text-deviating information for hierarchical reasoning.

3.3 Hierarchical Evidence Reasoning

The interpreted evidence bank contains heterogeneous cues across modalities, views, and semantic states. Direct pooling may obscure individual evidence and their interactions. We therefore organize the evidence through local, relational, and global reasoning.

Local Evidence Modeling. We employ N_L learnable queries to retrieve fine-grained evidence from the interpreted evidence bank:

$$H^L = \mathcal{A}_L\left(Q^L, \widetilde{E}\right), \tag{11}$$

where $\mathcal{A}_L$ denotes query-based cross-attention with contextual refinement, and $H^L \in \mathbb{R}^{N_L \times d}$ contains the resulting local evidence tokens.

Relational Evidence Modeling. To model interactions among local cues, we construct a continuous relation graph with $s_{ij} = f_r\big(\tanh(W_q h_i^L + W_k h_j^L)\big)$ and relational context $r_i = \sum_j \text{softmax}_j(s_{ij}) W_v h_j^L$. Fusing each token with its relational context gives $\overline{H}^L = \text{Fuse}(H^L, R)$, where $R = [r_1; \ldots; r_{N_L}]$. A smaller set of queries then abstracts the interaction-aware evidence:

$$H^R = \mathcal{A}_R\left(Q^R, \overline{H}^L\right), \tag{12}$$

where $H^R \in \mathbb{R}^{N_R \times d}$ and $N_R < N_L$.

Global Evidence Abstraction. Global queries further aggregate local and relational evidence:

$$H^G = \mathcal{A}_G\left(Q^G, \text{Concat}_{\text{tok}}(H^L, H^R)\right), \tag{13}$$

where $H^G \in \mathbb{R}^{N_G \times d}$ and $N_G < N_R < N_L$. The decreasing query counts progressively organize evidence from individual cues to interactions and global abstractions while retaining all three levels for pseudo-token projection.

To preserve sentiment-discriminative information across the hierarchy, we attach a prediction head to each level: $z^s = g_s\Big(\frac{1}{N_s}\sum_{i=1}^{N_s} h_i^s\Big)$, $s \in \{L, R, G\}$. The hierarchical prediction loss is:

$$\mathcal{L}_{\text{hier}} = \frac{1}{3}\sum_{s \in \{L,R,G\}} \ell_{\text{task}}(z^s, y), \tag{14}$$

where ℓ_{task} is MAE for regression and cross-entropy for classification. We further encourage local and relational predictions to approach the global prediction:

$$\mathcal{L}_{\text{prog}} = \begin{cases} \frac{1}{2} \sum_{s \in \{L,R\}} \left(z^s - \text{sg}(z^G)\right)^2, & \text{regression,} \\ \frac{1}{2} \sum_{s \in \{L,R\}} D_{\text{KL}}\left(\text{sg}(p^G) \,\|\, p^s\right), & \text{classification,} \end{cases} \tag{15}$$

where $p^s = \text{softmax}(z^s)$ for classification and $\text{sg}(\cdot)$ denotes stop-gradient. The global prediction remains trainable through $\mathcal{L}_{\text{hier}}$, while serving as a detached consistency target.

Incremental Pseudo-Token Projection. To avoid repeatedly introducing information shared across reasoning levels, we obtain compact summaries $U^s \in \mathbb{R}^{M_s \times d}$, $s \in \{L, R, G\}$, using separate query-based projectors, where M_L, M_R, and M_G denote the numbers of pseudo-token slots allocated to the local, relational, and global levels, respectively. We retain U^L and construct relational and global increments as:

$$\begin{aligned} \Delta U^R &= U^R - \text{MHA}(U^R, U^L, U^L), \\ \Delta U^G &= U^G - \text{MHA}(U^G, U^{<G}, U^{<G}), \end{aligned} \tag{16}$$

where $U^{<G} = \text{Concat}_{\text{tok}}(U^L, U^R)$. The resulting hierarchy-aware pseudo-tokens are:

$$P = \text{Concat}_{\text{tok}}\left(\psi(U^L) + s_L,\ \psi(\Delta U^R) + s_R,\ \psi(\Delta U^G) + s_G\right), \tag{17}$$

where ψ projects into the LLM embedding space and s_L, s_R, and s_G are learnable level embeddings.

Sentiment Prediction and Training Objective. We prepend P to the textual embeddings and feed the sequence into a frozen LLM for sentiment prediction. The trainable modules are optimized by:

$$\mathcal{L} = \mathcal{L}_{\text{LLM}} + \lambda_{\text{disc}}\mathcal{L}_{\text{disc}} + \lambda_{\text{hier}}\mathcal{L}_{\text{hier}} + \lambda_{\text{prog}}\mathcal{L}_{\text{prog}}, \tag{18}$$

where $\mathcal{L}_{\text{LLM}}$ is the primary sentiment prediction loss and the remaining terms provide discovery, hierarchical, and progressive consistency supervision.

4 Experiments

Datasets and Evaluation Metrics. We evaluate HEAR on four multimodal datasets: CMU-MOSEI (Zadeh et al., 2018), CH-SIMS-V2 (Liu et al., 2022), MELD (Poria et al., 2019), and CHERMA (Sun et al., 2023). For CMU-MOSEI, we report Acc-2, F1, Acc-7, MAE, and Corr, with Acc-2/F1 following the non-negative/negative criterion. For CH-SIMS-V2, we report Acc-2, F1, Acc2-weak, MAE, and Corr, where Acc-2/F1 use the non-positive/positive criterion and Acc2-weak evaluates samples within $[-0.4, 0.4]$. For MELD and CHERMA, we report accuracy (Acc) and weighted F1 (WF1). Further details are provided in Appendix A.1 and Appendix A.5.

Implementation Details. We use Qwen-1.8B (Bai et al., 2023), LLaMA2-7B (Touvron et al., 2023), and ChatGLM3-6B (GLM Team et al., 2024) as frozen LLM backbones, with non-textual features extracted following prior work (Yang et al., 2025b). All experiments are conducted on NVIDIA RTX 4090 GPUs over five random seeds, and average results are reported. All non-LLM modules are jointly optimized with AdamW, with hyperparameters selected on the validation set. Further implementation details are provided in Appendix A.3.

4.1 Comparison with State-of-the-Art Methods.

Baselines. We compare HEAR with conventional multimodal methods (TFN, LMF, MulT, MAG-BERT, MISA, Self-MM, MMIM, AV-MC, and CHFN), pretrained-model-based and LLM-based methods (UniMSE and UniSA, and MSE). For a controlled comparison, HEAR and MSE use the same Qwen-1.8B, LLaMA2-7B, and ChatGLM3-6B backbones. Results are reported in Table 1.

Comparison with conventional multimodal methods. HEAR outperforms conventional multimodal methods across all four benchmarks. HEAR-ChatGLM3-6B achieves 87.29 Acc-2 and 87.06 F1 on CMU-MOSEI, compared with 83.70 and 83.90 for CHFN, and improves Acc-2 and Acc2-w over AV-MC by 2.70 and 3.31 points on CH-SIMS-V2. It also achieves the best Acc and WF1 on MELD and CHERMA, showing effectiveness across sentiment prediction and emotion recognition.

Table 1: Comparison of sentiment prediction and emotion recognition results on CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA. The best and second-best results are highlighted in bold and underlined, respectively. The HEAR-ChatGLM3-6B row is shaded in light blue.

Model	CMU-MOSEI					CH-SIMS-V2					MELD		CHERMA	
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc2-w	MAE	Corr	Acc	WF1	Acc	WF1
TFN (Zadeh et al., 2017)	78.50	78.96	51.60	0.573	0.714	76.51	76.31	66.27	0.323	0.667	60.77	57.74	–	68.37
LMF (Liu et al., 2018)	80.54	80.94	51.59	0.576	0.717	77.05	77.02	69.34	0.343	0.638	61.15	58.30	–	68.23
MulT (Tsai et al., 2019)	81.15	81.56	52.84	0.559	0.733	79.50	79.59	69.61	0.317	0.703	–	–	–	69.24
MAG-BERT (Rahman et al., 2020)	82.51	82.77	50.41	0.583	0.741	79.79	79.78	71.87	0.334	0.691	–	–	–	–
MISA (Hazarika et al., 2020)	83.60	83.80	52.20	0.555	0.756	80.53	80.63	70.50	0.314	0.725	–	–	–	–
Self-MM (Yu et al., 2021)	82.81	82.53	53.46	0.530	0.765	79.01	78.89	71.87	0.335	0.640	–	–	–	–
MMIM (Han et al., 2021)	82.24	82.66	54.24	0.526	0.772	80.95	80.97	72.28	0.316	0.707	–	–	–	–
AV-MC (Liu et al., 2022)	–	–	–	–	–	82.50	82.55	74.54	0.297	0.732	–	–	–	–
CHFN (Guo et al., 2022)	83.70	83.90	54.30	0.525	0.778	–	–	–	–	–	–	–	–	–
UniMSE (Hu et al., 2022b)	85.86	85.79	54.39	0.523	0.773	–	–	–	–	–	65.09	65.51	–	–
UniSA$_{GPT2}$ (Li et al., 2023c)	71.02	–	41.36	0.838	–	–	–	–	–	–	48.12	31.26	–	–
UniSA$_{T5}$ (Li et al., 2023c)	84.22	–	52.50	0.546	–	–	–	–	–	–	64.52	62.17	–	–
UniSA$_{BART}$ (Li et al., 2023c)	84.93	–	50.03	0.587	–	–	–	–	–	–	62.34	62.22	–	–
MSE-Qwen-1.8B (Yang et al., 2025b)	84.12	83.45	52.02	0.558	0.725	80.44	80.24	73.09	0.311	0.678	62.18	59.87	70.38	70.21
MSE-LLaMA2-7B (Yang et al., 2025b)	86.74	86.51	55.57	0.501	0.787	75.53	75.44	68.61	0.382	0.553	65.14	63.66	71.58	71.41
MSE-ChatGLM3-6B (Yang et al., 2025b)	86.91	86.77	54.56	0.515	0.783	83.77	83.76	75.24	0.296	0.720	66.23	65.13	72.90	72.73
HEAR-Qwen-1.8B	85.77	85.39	53.38	0.539	0.757	81.55	81.46	74.41	0.292	0.701	63.90	62.04	73.82	73.85
HEAR-LLaMA2-7B	86.91	86.76	**55.96**	0.496	0.798	77.74	77.36	71.18	0.326	0.619	66.59	64.91	72.88	72.91
HEAR-ChatGLM3-6B	**87.29**	**87.06**	55.60	**0.487**	**0.809**	**85.20**	**85.21**	**77.85**	**0.261**	**0.759**	**67.23**	**66.31**	**74.63**	**74.66**

Comparison with trained-model-based and LLM-based methods. HEAR outperforms UniMSE and UniSA, with its variants achieving strong results across the evaluation metrics. More importantly, HEAR consistently surpasses MSE under the same three LLM backbones. With ChatGLM3-6B, HEAR reduces MAE from 0.515 to 0.487 on CMU-MOSEI and from 0.296 to 0.261 on CH-SIMS-V2, while improving WF1 from 65.13 to 66.31 on MELD and from 72.73 to 74.66 on CHERMA. These matched-backbone comparisons indicate that the gains cannot be attributed solely to the choice of LLM backbone and support the effectiveness of HEAR's structured evidence modeling and hierarchical reasoning.

4.2 Ablation Studies

We conduct ablation experiments on four benchmarks to evaluate the contributions of HEAR's core modules and their internal designs, with results reported in Table 2.

Module Ablation. Removing any core module consistently degrades performance, confirming their complementary contributions. Hierarchical reasoning has a particularly pronounced impact, with its removal reducing CMU-MOSEI Acc-7 from 55.60% to 53.20% and CH-SIMS-V2 Acc-2 from 85.20% to 82.79%. Evidence discovery is also critical: removing it reduces CH-SIMS-V2 Acc-2 from 85.20% to 83.13% and Corr from 0.759 to 0.744. Likewise, removing evidence interpretation leads to consistent performance degradation across all four benchmarks. These results validate the three-stage design of HEAR, where evidence discovery and interpretation provide structured multimodal cues for subsequent hierarchical reasoning.

Evidence Discovery. We further examine the evidence discovery module by retaining only individual types of discovered evidence. Neither temporal nor spectral evidence alone matches their joint use in Full HEAR. This difference is particularly evident on CH-SIMS-V2, where the Temporal Only and Spectral Only variants achieve Acc-2 scores of 83.72% and 81.02%, respectively, compared with 85.20% for Full HEAR. The Spectral Only variant also achieves only 72.46% Acc2-w, substantially below the 77.85% of the full model. Similar trends are observed for shared and private evidence across the benchmarks, demonstrating the complementary roles of temporal–spectral cues and shared–private representations in evidence discovery.

Evidence Interpretation. Modeling aligned or deviating evidence alone consistently underperforms their joint use. On CMU-MOSEI, the Aligned Only and Deviating Only variants obtain Acc-7 scores of 54.73% and 54.30%, respectively, compared with 55.60% for Full HEAR. On CH-SIMS-V2, their Acc-2 scores are 83.96% and 84.37%, respectively, compared with 85.20% for the full

Table 2: Ablation study of HEAR on four benchmark datasets. All variants use the same LLM backbone. ↑ and ↓ indicate that higher and lower values are better, respectively.

Variant	CMU-MOSEI					CH-SIMS-V2					MELD		CHERMA	
	Acc-2↑	F1↑	Acc-7↑	MAE↓	Corr↑	Acc-2↑	F1↑	Acc2-w↑	MAE↓	Corr↑	Acc↑	WF1↑	Acc↑	WF1↑
Module Ablation														
w/o Evidence Discovery	87.03	86.67	55.55	0.488	0.808	83.13	83.03	75.65	0.264	0.744	66.99	66.18	74.56	74.62
w/o Evidence Interpretation	87.18	86.99	55.51	0.489	0.807	83.50	83.38	76.11	0.266	0.757	67.05	66.03	73.89	74.08
w/o Hierarchical Reasoning	86.81	86.50	53.20	0.526	0.788	82.79	82.66	75.28	0.269	0.741	65.89	65.25	74.57	74.36
Evidence Discovery														
Temporal Only	86.94	86.76	54.85	0.493	0.806	83.72	83.70	76.07	0.266	0.752	67.21	66.24	74.30	74.29
Spectral Only	87.16	86.99	55.51	0.488	0.808	81.02	80.72	72.46	0.287	0.720	65.02	64.73	72.41	72.51
Shared Only	86.90	86.76	55.31	0.491	0.808	83.93	83.77	76.77	0.264	0.754	66.93	66.30	73.52	73.58
Private Only	87.12	86.82	55.48	0.492	0.806	84.02	84.01	76.15	0.262	0.758	67.01	66.05	73.54	73.43
Evidence Interpretation														
Aligned Only	87.12	86.84	54.73	0.493	0.804	83.96	83.85	76.77	0.265	0.750	66.48	66.37	74.48	74.53
Deviating Only	87.16	86.98	54.30	0.498	0.807	84.37	84.32	76.81	0.264	0.757	67.17	66.28	74.43	74.51
Hierarchical Reasoning														
Local Only	86.76	86.65	55.08	0.496	0.805	82.30	82.04	75.36	0.277	0.732	66.02	65.40	73.26	73.49
Local + Relational	87.19	86.92	55.48	0.508	0.806	84.14	84.02	77.02	0.272	0.745	66.86	66.04	74.36	74.47
Local + Global	87.08	86.55	54.97	0.490	0.799	84.04	84.04	75.98	0.277	0.748	66.13	65.44	74.30	73.85
Full HEAR	**87.29**	**87.06**	**55.60**	**0.487**	**0.809**	**85.20**	**85.21**	**77.85**	**0.261**	**0.759**	**67.23**	**66.31**	**74.63**	**74.66**

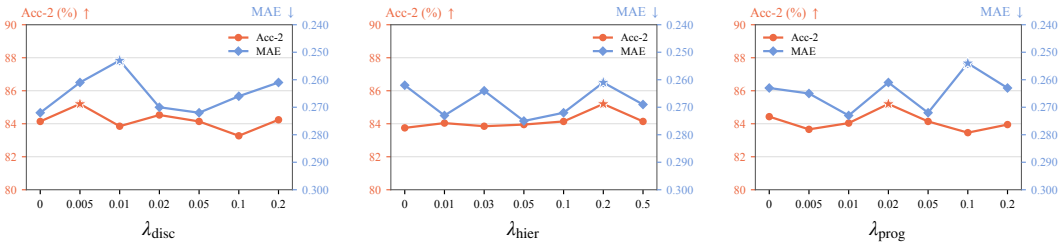


Figure 3: Sensitivity analysis of the discovery, hierarchical supervision, and progressive consistency loss weights on Acc-2 and MAE.

model. Similar trends are observed on MELD and CHERMA. These results suggest that aligned evidence captures cross-modal agreement, whereas deviating evidence preserves informative discrepancies among modalities. Jointly interpreting both types therefore enables HEAR to exploit a more comprehensive set of cross-modal relationships.

Hierarchical Reasoning. We examine the progressive reasoning hierarchy from local evidence understanding to relational reasoning and global integration. As the reasoning hierarchy becomes more complete, clear performance improvements are observed. On CH-SIMS-V2, adding relational reasoning improves Acc-2 from 82.30% to 84.14% and Acc2-w from 75.36% to 77.02%, while the complete hierarchy further improves them to 85.20% and 77.85%, respectively. Moreover, bypassing relational reasoning with Local+Global generally yields inferior performance, highlighting the importance of modeling inter-evidence dependencies before global integration. These results support the complementary roles of local, relational, and global reasoning in HEAR.

Loss Weight Sensitivity. We analyze the sensitivity of three loss weights on CH-SIMS-V2 in Fig. 3. For λ_{disc}, the best Acc-2 and MAE are achieved at 0.005 and 0.01, respectively, while $\lambda_{\text{hier}} = 0.2$ performs best on both metrics. For λ_{prog}, 0.02 yields the highest Acc-2, whereas 0.1 achieves the lowest MAE, indicating a trade-off between classification and regression objectives. Overall, the non-monotonic trends suggest that moderate auxiliary supervision is preferable to overly strong constraints. Sensitivity analyses of other hyperparameters are provided in Appendix B.5.

4.3 Parameter Efficiency

Figure 4 compares HEAR across four datasets and three LLM backbones. HEAR updates only 0.28%–0.94% of the parameters, leaving over 99% frozen. ChatGLM achieves the highest accuracy across all datasets with only 0.31%–0.33% trainable parameters, while Qwen requires the fewest

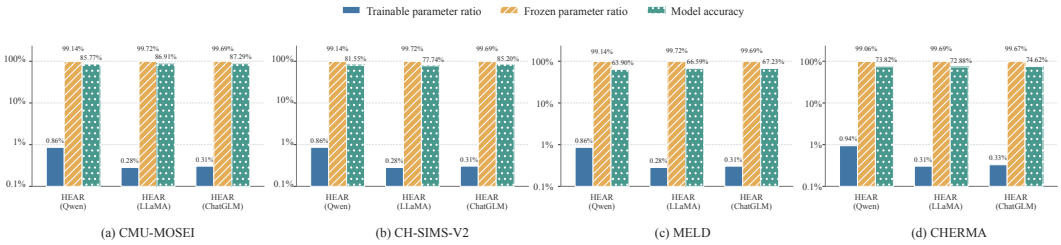


Figure 4: Trainable and frozen parameter ratios and accuracy of HEAR across four datasets and three language model backbones.

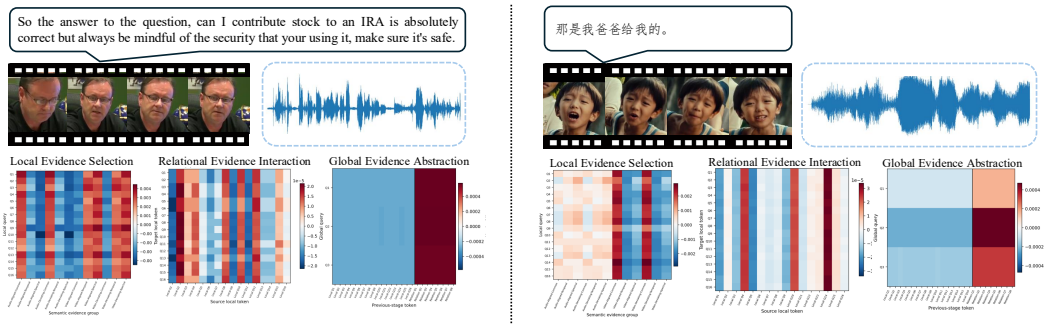


Figure 5: Visualization of hierarchical evidence reasoning in HEAR. Two examples illustrate cross-modal discrepancies (a, CMU-MOSEI) and consistency (b, CH-SIMS-V2) through local evidence selection, relational interaction, and global abstraction.

trainable parameters in absolute terms. Notably, LLaMA outperforms Qwen on CMU-MOSEI and MELD but not on CH-SIMS-V2 and CHERMA, suggesting that backbone scale does not consistently translate into better performance. Overall, HEAR remains parameter-efficient across backbones, with performance depending more on backbone and task compatibility than model scale.

4.4 Visualization of Hierarchical Evidence Reasoning

Visualization Analysis. Figure 5 illustrates HEAR's hierarchical evidence reasoning under two multimodal scenarios. In CMU-MOSEI (Fig. 5a), neutral text coexists with negative facial cues and relatively positive vocal characteristics, creating cross-modal discrepancies. HEAR progressively models local evidence and their relations, allowing the relational stage to capture cross-modal interactions before global integration, yielding a prediction of $+0.434$ close to the ground truth of $+0.333$. In CH-SIMS-V2 (Fig. 5b), consistent negative facial and vocal cues are progressively aggregated, producing a prediction of -0.960 that closely matches the ground truth of -1.000. Together, these examples illustrate how HEAR transitions from local evidence understanding to relational interaction and global integration under both discrepant and consistent conditions.

5 Conclusion

We presented HEAR, a hierarchical evidence-aware reasoning framework that shifts multimodal sentiment analysis from representation fusion toward structured evidence reasoning. HEAR preserves fine-grained multimodal cues and informative cross-modal discrepancies through text-conditioned evidence interpretation, and progressively organizes them through local, relational, and global reasoning before integration with a frozen LLM. Experiments on four benchmarks demonstrate consistent improvements over strong fusion-based and LLM-based baselines, supporting the effectiveness of explicitly discovering, contextualizing, and reasoning over sentiment evidence for multimodal sentiment understanding.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the design of research experiments and the translation of text. We did not use generative AI tools for other tasks requiring mandatory disclosure under the ICLR 2027 AI Policy for Authors, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to assist with modifying and debugging code, summarizing and analyzing existing literature, improving the clarity and readability of the manuscript, and suggesting keywords for the paper. All AI-assisted content was reviewed and verified by the authors. AI-assisted experimental suggestions were critically evaluated by the authors before being incorporated into the experimental design. AI-assisted code modifications and debugging suggestions were manually inspected and tested for correctness. Literature summaries and analyses were checked by the authors against the original sources, and all language edits, translations, and suggested keywords were reviewed to ensure that they accurately reflected the authors' intended meaning and the content of the work. We take responsibility for the final content of this work, including text, claims, code, experimental design, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work relies exclusively on existing benchmark datasets, including CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA. We did not collect any new human-subject data, and all datasets were used in accordance with their respective terms of use. Our work focuses on multimodal sentiment analysis and does not attempt to identify individual participants. We acknowledge the potential privacy and ethical concerns associated with the misuse of multimodal sentiment analysis, such as surveillance or profiling, and encourage the responsible use of the proposed method. We adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we will release the implementation of our proposed method, including data preprocessing, model training, and evaluation code. We provide detailed descriptions of the experimental setup, including dataset preprocessing, model architecture, training procedures, and hyperparameter settings, in the main paper and appendix. We also report the implementation details and evaluation protocols used for all benchmark datasets. These resources are intended to enable the reproduction of the experimental results reported in this paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang, and Yanbin Hao. Mixture of multimodal adapters for sentiment analysis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1822–1833, 2025.

Hongju Cheng, Zizhen Yang, Xiaoqi Zhang, and Yang Yang. Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion. *IEEE Transactions on Affective Computing*, 14(4):3149–3163, 2023. doi: 10.1109/TAFFC.2023.3265653.

Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):270:1–270:38, 2023. doi: 10.1145/3586075.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. MAGMA – multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2416–2428, 2022.

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.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3394–3402, 2022.

Wei Han, Hui Chen, and Soujanya Poria. 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*, pp. 9180–9192, 2021.

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, 2020.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7037–7041, 2022a.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7837–7851, 2022b.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021.

Peiyuan Jiang, Yao Liu, Qiao Liu, Zongshun Zhang, Jiaye Yang, Lu Liu, and Daibing Yao. DRKF: decoupled representations with knowledge fusion for multimodal emotion recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 955–964, 2025a.

Shenjie Jiang, Zhuoyu Wang, Xuecheng Wu, Hongru Ji, Mingxin Li, Xianghua Li, and Chao Gao. DDSE: a decoupled dual-stream enhanced framework for multimodal sentiment analysis with text-centric SSM. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5893–5902, 2025b.

Xin Kang, Xuefeng Shi, Yunong Wu, and Fuji Ren. Active learning with complementary sampling for instructing class-biased multi-label text emotion classification. *IEEE Transactions on Affective Computing*, 14:523–536, 2023.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. GA2MIF: graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Transactions on Affective Computing*, 15(1):130–143, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023a.

Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled Mamba: enhanced multimodal fusion with coupled state space model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59808–59832, 2024b.

Yan Li, Xiangyuan Lan, Haifeng Chen, Ke Lu, and Dongmei Jiang. Multimodal PEAR chain-of-thought reasoning for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(9):286:1–286:23, 2024c. doi: 10.1145/3672398.

Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. AlignMamba: enhancing multimodal Mamba with local and global cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24774–24784, 2025.

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, 2023b.

Yong Li, Yuanzhi Wang, Yi Ding, Shiqing Zhang, Ke Lu, and Cuntai Guan. Decoupled hierarchical distillation for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6912–6928, 2026. doi: 10.1109/TPAMI.2026.3660754.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. EmoCaps: emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1610–1618, 2022.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. UniSA: unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6132–6142, 2023c. doi: 10.1145/3581783.3612336.

Zongxi Li, Haoran Xie, Gary Cheng, and Qing Li. Word-level emotion distribution with two schemas for short text emotion classification. *Knowledge-Based Systems*, 227:107163, 2021.

Sheng Liang, Mengjie Zhao, and Hinrich Schuetze. Modular and parameter-efficient multimodal fusion with prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2976–2985, 2022.

Tian Liang, Jing Huang, Ming Kong, Luyuan Chen, and Qiang Zhu. Querying as prompt: parameter-efficient learning for multimodal language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26845–26855, 2024.

Xixun Lin, Rui Liu, Yanan Cao, Lixin Zou, Qian Li, Yongxuan Wu, Yang Liu, Dawei Yin, and Guandong Xu. Contrastive modality-disentangled learning for multimodal recommendation. *ACM Transactions on Information Systems*, 43(3):70:1–70:31, 2025. doi: 10.1145/3715876.

Xinjing Liu, Pengyue Lin, Xinyu Tu, Wenqi Jia, Chen Jiang, and Ruifan Li. SDG-MLLM: injecting structured dialogue graphs into MLLM for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 14114–14121, 2025. doi: 10.1145/3746027.3762071.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 247–258, 2022.

Zhaowei Liu, Sheng Liu, Weiqing Yan, Peng Song, Yongchao Song, and Rufei Gao. ATGFB-MFF: adaptive text-guided fiber bundle feature fusion with LLMs for multimodal sentiment analysis and emotion recognition in conversations. In *Proceedings of the ACM Web Conference 2026*, pp. 7442–7453, 2026. doi: 10.1145/3774904.3792577.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher 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, 2018.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2554–2562, 2021.

Xiongjian Lv, Yimin Wen, and Hang Yu. DiffuFuse: diffusion-driven dual-stream fusion framework for multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8458–8467, 2025.

Lisong Ou and Zhixin Li. TPFN: a text-guided progressive fusion network for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 17(2):1585–1597, 2026. doi: 10.1109/TAFFC.2026.3663280.

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, 2019.

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, 2020.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 658–670, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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, 2019.

Sheng Wu, Dongxiao He, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Enriching multimodal sentiment analysis through textual emotional descriptions of visual-audio content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1601–1609, 2025.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1642–1651, 2022.

Hao Yang, Yanyan Zhao, Yang Wu, Shilong Wang, Tian Zheng, Hongbo Zhang, Zongyang Ma, Wanxiang Che, Shijin Wang, Si Wei, and Bing Qin. Large language models meet text-centric multimodal sentiment analysis: a survey. *Science China Information Sciences*, 68(10):200101, 2025a. doi: 10.1007/s11432-024-4593-8.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. ConFEDE: contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7617–7630, 2023. doi: 10.18653/v1/2023.acl-long.421.

Yang Yang, Xunde Dong, and Yupeng Qiang. MSE-Adapter: a lightweight plugin endowing LLMs with the capability to perform multimodal sentiment analysis and emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25642–25650, 2025b.

Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, et al. Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms. *Advances in Neural Information Processing Systems*, 36:52692–52704, 2023a.

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

Yakun Yu, Mingjun Zhao, Shi-ang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. ConKI: contrastive knowledge injection for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13610–13624, 2023b.

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, 2017. doi: 10.18653/v1/D17-1115.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, 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, 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, 2023.

Xianbing Zhao, Shengzun Yang, and Buzhou Tang. CCAF: coarse-to-fine cross-modal alignment and fusion for multimodal sentiment analysis. In *Proceedings of the ACM Web Conference 2026*, pp. 7421–7430, 2026. doi: 10.1145/3774904.3792569.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. GLoMo: global-local modal fusion for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1800–1809, 2024.

APPENDIX

To provide further details on our method and experiments, we organize the appendix as follows:

- **Appendix A** presents experimental details, including dataset descriptions, implementation settings, hyperparameters, evaluation metrics, and dataset-specific prompts.
- **Appendix B** provides baseline descriptions, extended quantitative analyses, additional ablation studies, and hyperparameter sensitivity analyses.
- **Appendix C** presents qualitative visualizations and case studies of hierarchical evidence reasoning across the four datasets.

A EXPERIMENTAL DETAILS

A.1 MOTIVATION FOR DATASET SELECTION

To comprehensively evaluate the effectiveness and generalizability of HEAR, we select four multimodal benchmarks, namely CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA, covering both multimodal sentiment analysis and emotion recognition. Specifically, CMU-MOSEI and CH-SIMS-V2 evaluate sentiment under English and Chinese contexts, respectively, while MELD and CHERMA extend the evaluation to categorical emotion recognition. Such task and linguistic diversity is particularly relevant to HEAR, as our framework is designed to reason over sentiment-relevant evidence rather than relying on task-specific multimodal fusion patterns.

More importantly, these benchmarks provide diverse settings for examining the central motivation of HEAR: perceptually salient multimodal cues are not always sentiment-relevant, and different modalities may provide consistent or discrepant evidence for sentiment inference. By conditioning fine-grained acoustic and visual evidence on textual semantics, HEAR explicitly preserves both text-aligned and text-deviating evidence and progressively reasons over them at multiple levels. Evaluating across different tasks, languages, and emotional label spaces therefore allows us to examine whether such evidence-aware reasoning remains effective beyond a particular dataset or prediction setting. Collectively, these four benchmarks provide a diverse testbed for assessing the effectiveness, robustness, and generalizability of HEAR. Table 3 summarizes their key characteristics.

Table 3: Statistics and characteristics of the four datasets used in our experiments.

Dataset	Language	Task	Label Space	# Samples
CMU-MOSEI	English	Sentiment Analysis	$[-3, 3]$	22,856
CH-SIMS-V2	Chinese	Sentiment Analysis	$[-1, 1]$	4,403
MELD	English	Emotion Recognition	7 classes	13,708
CHERMA	Chinese	Emotion Recognition	7 classes	28,717

A.2 DETAILED DATASET DESCRIPTION

We provide detailed descriptions of the four benchmark datasets used in our experiments, including two multimodal sentiment analysis datasets (CMU-MOSEI and CH-SIMS-V2) and two multimodal emotion recognition datasets (MELD and CHERMA).

CMU-MOSEI. CMU-MOSEI (Zadeh et al., 2018) is an English multimodal sentiment analysis dataset comprising 22,856 video clips collected from YouTube. Each clip is annotated with a sentiment score ranging from -3 (strongly negative) to $+3$ (strongly positive). Following the standard data split, 16,326 clips are used for training, 1,871 for validation, and 4,659 for testing.

CH-SIMS-V2. CH-SIMS-V2 (Liu et al., 2022) is a Chinese multimodal sentiment analysis dataset consisting of 4,403 curated video clips collected from movies, TV shows, and variety programs. Each clip is manually annotated with a sentiment score ranging from -1 (strongly negative) to $+1$ (strongly positive). Following the standard split, the dataset contains 2,722 training, 647 validation, and 1,034 test samples.

MELD. MELD (Poria et al., 2019) is an English multimodal dataset for emotion recognition in conversational settings. It contains 1,433 dialogues and 13,708 utterances collected from the TV series *Friends*. Each utterance is annotated with one of seven emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise.

CHERMA. CHERMA (Sun et al., 2023) is a Chinese multimodal dataset for emotion recognition in conversational settings, comprising 28,717 utterances collected from TV dramas, variety shows, and films. Each utterance is annotated with one of seven emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise, consistent with the emotion label space adopted by MELD.

A.3 Implementation Details

We adopt Qwen-1.8B (Bai et al., 2023), LLaMA2-7B (Touvron et al., 2023), and ChatGLM3-6B (GLM Team et al., 2024) as the LLM backbones. The LLM parameters remain frozen throughout training, while all non-LLM components of HEAR are jointly optimized. Following prior work (Yang et al., 2025b), we use pre-extracted acoustic and visual features as the non-textual inputs.

All experiments are conducted on NVIDIA RTX 4090 GPUs. We run each experiment with five random seeds, $\{1111, 2222, 3333, 4444, 5555\}$, and report the average performance over the five runs. We use AdamW as the optimizer and apply a cosine learning-rate schedule with a linear warm-up over the first 6% of the total training steps. The learning rate is set to 1×10^{-4} for the evidence discovery, evidence interpretation, and hierarchical reasoning modules, and 5×10^{-5} for the projection module. We use a weight decay of 0.01, excluding bias and one-dimensional parameters, and clip the gradient norm to 1.0. The batch sizes are set to 8, 8, 6, and 4 for CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA, respectively. We employ early stopping with a patience of 5 epochs and select the best checkpoint according to the validation performance. Specifically, MAE is used for CMU-MOSEI and CH-SIMS-V2, whereas weighted F1 is used for MELD and CHERMA. We provide additional implementation details, including training settings and hyperparameters.

A.4 Hyperparameter Settings

Table 4 summarizes the hyperparameter settings used for all four datasets. Unless otherwise specified, we keep the model architecture and optimization-related hyperparameters fixed across datasets. In particular, the hidden dimension is set to 256, with four attention heads and a dropout rate of 0.1. For hierarchical evidence reasoning, the numbers of queries at the local, relational, and global levels are set to $(N_L, N_R, N_G) = (16, 8, 3)$, respectively. The corresponding numbers of projected pseudo-tokens are set to $(K_L, K_R, K_G) = (4, 3, 2)$. We use 32 frequency bins, and each evidence tokenizer produces four tokens per branch.

For optimization, the hierarchical, discovery, and progressive loss weights are fixed to $\lambda_{\text{hier}} = 0.2$, $\lambda_{\text{disc}} = 0.005$, and $\lambda_{\text{prog}} = 0.02$, respectively, across all datasets. Most training hyperparameters are also shared, while the batch size and learning-rate schedule horizon are adjusted according to the dataset. Specifically, the batch sizes are set to 8, 8, 6, and 4 for CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA, respectively. For checkpoint selection, we use validation MAE for the sentiment analysis datasets (CMU-MOSEI and CH-SIMS-V2) and validation weighted F1 for the emotion recognition datasets (MELD and CHERMA).

A.5 Evaluation Metrics

We evaluate our framework on CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA using regression and classification metrics. For sentiment analysis, we use MAE, Corr, and discretized sentiment classification metrics. For emotion recognition on MELD and CHERMA, we use accuracy (Acc) and weighted F1 (WF1). The metrics are described below.

MAE. Mean absolute error measures the average absolute difference between predicted and ground-truth sentiment scores. Lower MAE indicates more accurate predictions.

Table 4: Hyperparameter settings of HEAR on the four benchmark datasets.

Hyperparameter	CMU-MOSEI	CH-SIMS-V2	MELD	CHERMA
Training Settings				
Batch size	8	8	6	4
Encoder learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Projection learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
Early-stopping patience	5	5	5	5
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight decay	0.01	0.01	0.01	0.01
LR scheduler	Cosine	Cosine	Cosine	Cosine
LR warmup ratio	0.06	0.06	0.06	0.06
Checkpoint selection metric	MAE ↓	MAE ↓	Weighted F1 ↑	Weighted F1 ↑
Model Settings				
Hidden dimension d	256	256	256	256
LLM embedding dimension	4096	4096	4096	4096
Attention heads	4	4	4	4
Dropout	0.1	0.1	0.1	0.1
Evidence tokens (per view)	4	4	4	4
Frequency bins	32	32	32	32
Hierarchical queries (N_L, N_R, N_G)	$(16, 8, 3)$	$(16, 8, 3)$	$(16, 8, 3)$	$(16, 8, 3)$
Pseudo-tokens (M_L, M_R, M_G)	$(4, 3, 2)$	$(4, 3, 2)$	$(4, 3, 2)$	$(4, 3, 2)$
Loss Settings				
Discovery loss weight λ_{disc}	0.005	0.005	0.005	0.005
Hierarchical loss weight λ_{hier}	0.2	0.2	0.2	0.2
Progressive loss weight λ_{prog}	0.02	0.02	0.02	0.02

Corr. Pearson correlation measures the linear association between predicted and ground-truth sentiment scores:

$$\text{Corr} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{19}$$

where $\hat{y}_i$ and y_i denote predicted and ground-truth scores, and $\bar{\hat{y}}$ and $\bar{y}$ are their means. Higher Corr indicates stronger linear agreement. For CH-SIMS-V2, both scores are clipped to $[-1, 1]$ before evaluation; CMU-MOSEI regression metrics use unclipped scores.

Acc-2. Binary sentiment accuracy is the proportion of correctly predicted sentiment polarities. On CH-SIMS v2, scores greater than zero are positive, and the remaining scores are non-positive. On CMU-MOSEI, we evaluate both Has0 and Non0 settings: Has0 includes neutral references and treats zero as positive, whereas Non0 excludes neutral references and treats predicted zero as negative.

Acc2-weak. For CH-SIMS-V2, this metric measures binary sentiment accuracy on the subset whose ground-truth scores lie in $[-0.4, 0.4]$. It uses the same classification threshold as Acc-2 and assesses performance on weak sentiment expressions.

Acc-3, Acc-5, and Acc-7. These metrics measure the proportion of predictions assigned to the correct sentiment category after discretization. CH-SIMS-V2 uses threshold-based three-class and five-class partitions, whereas CMU-MOSEI uses five-class and seven-class labels obtained by clipping and rounding continuous scores.

Acc. For MELD and CHERMA, accuracy measures the proportion of correctly predicted labels across the seven emotion classes.

Prompt for CMU-MOSEI

Prompt: Please predict the sentiment intensity of the above multimodal content in the range [-3.0, +3.0].
Response: The sentiment is

Prompt for CH-SIMS-V2

Prompt:请对上述多模态内容的情感强度进行预测，范围在[-1.0, 1.0]之间。
响应：情感为

Prompt for MELD

Prompt: Please recognize the emotion of the above multimodal content from the target set <neutral:0, surprise:1, fear:2, sadness:3, joy:4, disgust:5, anger:6>.
Response: The emotion is

Prompt for CHERMA

Prompt:请选择适用于上述多模态内容的情绪标签：<愤怒:0, 厌恶:1, 恐惧:2, 高兴:3, 平静:4, 悲伤:5, 惊奇:6>。
响应：情感为

Figure 6: Dataset-specific prompts and response formats used for CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA.

F1 and WF1. The F1 score for class c combines precision P_c and recall R_c. Our implementation uses support-weighted averaging for both sentiment F1 and emotion WF1:

$$\mathrm{F1}_c = \frac{2P_cR_c}{P_c + R_c}, \qquad \mathrm{WF1} = \sum_{c \in \mathcal{C}} \frac{n_c}{N} \mathrm{F1}_c, \tag{20}$$

where n_c is the number of ground-truth samples in class c, and N is the total number of evaluated samples. Undefined per-class F1 scores are assigned zero. For sentiment analysis, the score is computed over the binary sentiment classes; for emotion recognition, it is computed over the seven emotion categories.

A.6 Dataset-Specific Prompts

To accommodate different task formulations, label spaces, and languages, we adopt dataset-specific prompts for each benchmark, as illustrated in Fig. 6. Specifically, CMU-MOSEI and CH-SIMS-V2 formulate sentiment analysis as intensity regression with dataset-specific score ranges, whereas MELD and CHERMA perform emotion classification over their respective label sets. We use English prompts for CMU-MOSEI and MELD and Chinese prompts for CH-SIMS-V2 and CHERMA, consistent with the primary language of each dataset. The exact prompts and response formats used in our experiments are provided in Fig. 6.

B Additional Experimental Results

Details of Baselines. We provide brief descriptions of the compared baselines, grouped into four categories as follows.

Traditional Fusion Methods. **TFN** (Zadeh et al., 2017): Captures unimodal and cross-modal interactions through tensor fusion over all combinations of modalities. **LMF** (Liu et al., 2018): Reduces

the computational cost of multimodal tensor fusion via low-rank decomposition. **MulT** (Tsai et al., 2019): Introduces cross-modal attention to directly model fine-grained interactions among modalities through temporal attention flows. **MAG-BERT** (Rahman et al., 2020): Integrates non-textual information into the BERT encoder through an adaptive gating mechanism for multimodal alignment. **CHFN** (Guo et al., 2022): Hierarchically fuses multiple modality pairs at different semantic levels for multimodal representation learning. **MMGCN** (Hu et al., 2021): Employs multimodal graph convolutional networks to learn multimodal sentiment representations. **PMR** (Lv et al., 2021): Employs a message hub and cross-modal attention to progressively exchange information and reinforce modality features for emotion recognition from unaligned sequences.

Representation Learning Methods. **MISA** (Hazarika et al., 2020): Explicitly learns modality-invariant and modality-specific representations to capture shared and private multimodal information. **Self-MM** (Yu et al., 2021): Generates unimodal supervision through self-supervised label generation and jointly learns shared and modality-specific representations with adaptive task weighting. **MMIM** (Han et al., 2021): Models multimodal interactions through hierarchical memory fusion and attention mechanisms. **MM-DFN** (Hu et al., 2022a): Decomposes multimodal features into shared and private representations and aligns them through contrastive objectives. **ConFEDE** (Yang et al., 2023): Decomposes multimodal representations into shared and modality-specific subspaces through contrastive learning for multimodal sentiment analysis. **ALMT** (Zhang et al., 2023): Performs language-guided cross-modal alignment through adaptive hyper-modality learning to enhance sentiment representations. **AV-MC** (Liu et al., 2022): Mixes acoustic and visual modalities across videos to augment multimodal contexts and strengthen non-verbal cues for sentiment prediction. **GA2MIF** (Li et al., 2024a): Enhances multimodal fusion through gated adaptive alignment and mutual-information-based fusion. **EmoCaps** (Li et al., 2022): Combines multimodal emotion vectors with sentence representations into emotion capsules and models conversational context for emotion recognition.

Recent Methods. **CMamba** (Li et al., 2024b): Couples modality-specific state chains through cross-modal hidden-state transitions to capture intra- and inter-modal dependencies for multimodal fusion. **DDSE** (Jiang et al., 2025b): Decouples shared and private representations and uses text-guided Mamba for progressive cross-modal interactions that preserve sentiment cues and reduce redundancy. **EMOE** (Fang et al., 2025): Combines modality experts with sample-specific weighting and unimodal distillation to adaptively fuse modalities while preserving unimodal predictive capabilities. **AlignMamba** (Li et al., 2025): Integrates optimal transport for local token alignment and maximum mean discrepancy for global distribution alignment with Mamba-based multimodal fusion. **DiffuFuse** (Lv et al., 2025): Uses text-guided diffusion to separate shared and private semantics, dual-stream fusion to model their interactions, and orthogonal projection to reduce redundancy. **LFMIM** (Sun et al., 2023): Preserves modality independence through separately supervised unimodal transformers and fuses their representations with a multimodal transformer for emotion recognition. **DRKF** (Jiang et al., 2025a): Decouples shared and modality-specific representations and performs dominant-modality-guided fusion while preserving discriminative cues of emotional inconsistency.

LLM-based Methods. **UniMSE** (Hu et al., 2022b): Unifies multimodal sentiment analysis and emotion recognition in conversation within a generation-based framework using T5 as the backbone. **UniSA** (Li et al., 2023c): Formulates different multimodal sentiment analysis subtasks within a unified sequence-generation framework. It includes BART-, T5-, and GPT2-based variants (UniSA-BART, UniSA-T5, and UniSA-GPT2), which are pretrained with contrastive objectives and subsequently fine-tuned on sentiment datasets. **SPAE** (Yu et al., 2023a): Encodes multimodal inputs into token representations compatible with frozen LLMs using a semantic pyramid autoencoder. **MSE-Adapter** (Yang et al., 2025b): Integrates large language models with the MSE-Adapter architecture for text-guided multimodal generation.

B.1 Extended Analysis for CMU-MOSEI

Comparison with Conventional Multimodal Methods. HEAR consistently outperforms conventional fusion- and representation-based methods on CMU-MOSEI. Earlier approaches such as TFN, LMF, and MulT achieve Acc-2 below 82%, while more advanced methods, including MISA, MMIM, CHFN, and UniMSE, progressively improve multimodal modeling. Among them, UniMSE achieves strong performance with 85.86% Acc-2 and 85.79% F1. In comparison, HEAR-

ChatGLM3-6B further improves these results to 87.29% and 87.06%, respectively, while reducing MAE from 0.523 to 0.487. These results suggest that organizing multimodal information as structured evidence provides an effective alternative to directly learning a fused multimodal representation.

Comparison with LLM-based Methods. HEAR also consistently improves over MSE-Adapter under the same LLM backbones. With Qwen-1.8B, HEAR improves Acc-2 from 84.12% to 85.77% and Corr from 0.725 to 0.757. Similar improvements are observed with LLaMA2-7B, where HEAR achieves the best Acc-7 of 55.96% and improves Corr from 0.787 to 0.798. With ChatGLM3-6B, HEAR further increases Acc-2/F1 from 86.91%/86.77% to 87.29%/87.06%, while reducing MAE from 0.515 to 0.487. The consistent gains across different backbones indicate that the improvements arise not merely from the underlying LLM, but from HEAR's structured evidence modeling and hierarchical reasoning.

Overall Analysis. Overall, HEAR achieves the best Acc-2 (87.29%), F1 (87.06%), MAE (0.487), and Corr (0.809), while HEAR-LLaMA2-7B obtains the best Acc-7 (55.96%). The improvements span both discrete sentiment classification and continuous sentiment intensity estimation, suggesting that HEAR effectively preserves fine-grained sentiment evidence rather than relying primarily on globally dominant multimodal information. Together with the consistent improvements over the corresponding MSE-Adapter backbones, these results support the effectiveness of shifting multimodal sentiment modeling from direct representation fusion toward structured evidence reasoning.

Table 5: Comparison of sentiment prediction results on CMU-MOSEI. The best and second-best results are highlighted in bold and underlined, respectively. * denotes results quoted from prior work.

Model	Acc-2↑	F1↑	Acc-7↑	MAE↓	Corr↑
TFN* (Zadeh et al., 2017)	78.50	78.96	51.60	0.573	0.714
LMF* (Liu et al., 2018)	80.54	80.94	51.59	0.576	0.717
MulT* (Tsai et al., 2019)	81.15	81.56	52.84	0.559	0.733
MAG-BERT* (Rahman et al., 2020)	82.51	82.77	50.41	0.583	0.741
MISA (Hazarika et al., 2020)	83.60	83.80	52.20	0.555	0.756
Self-MM* (Han et al., 2021)	82.81	82.53	53.46	0.530	0.765
MMIM (Hu et al., 2022a)	82.24	82.66	54.24	0.526	0.772
CHFN (Guo et al., 2022)	83.70	83.90	54.30	0.525	0.778
UniMSE (Hu et al., 2022b)	85.86	85.79	54.39	0.523	0.773
UniSA$_{GPT2}$ (Li et al., 2023c)	71.02	–	41.36	0.838	–
UniSA$_{T5}$ (Li et al., 2023c)	84.22	–	52.50	0.546	–
UniSA$_{BART}$ (Li et al., 2023c)	84.93	–	50.03	0.587	–
MSE-Qwen-1.8B (Yang et al., 2025b)	84.12	83.45	52.02	0.558	0.725
MSE-LLaMA2-7B (Yang et al., 2025b)	86.74	86.51	<u>55.57</u>	0.501	0.787
MSE-ChatGLM3-6B (Yang et al., 2025b)	<u>86.91</u>	<u>86.77</u>	54.56	0.515	0.783
HEAR-Qwen-1.8B	85.77	85.39	53.38	0.539	0.757
HEAR-LLaMA2-7B	<u>86.91</u>	86.76	**55.96**	<u>0.496</u>	<u>0.798</u>
HEAR-ChatGLM3-6B	**87.29**	**87.06**	55.60	**0.487**	**0.809**

B.2 Extended Analysis for CH-SIMS-V2

Comparison with Conventional Multimodal Methods. HEAR-ChatGLM3-6B outperforms the conventional multimodal baselines on all metrics for which comparison results are available on CH-SIMS-V2. Among these baselines, CMamba achieves the highest Acc-2 and F1 of 83.40% and 83.50%, respectively, together with the lowest MAE of 0.287 and the highest Corr of 0.758. HEAR-ChatGLM3-6B improves Acc-2 and F1 to 85.20% and 85.21%, representing gains of 1.80 and 1.71 percentage points, respectively, while reducing MAE to 0.261. Its Corr of 0.759 is marginally higher than that of CMamba. For weak sentiment expressions, HEAR-ChatGLM3-6B achieves 77.85% Acc2-w, exceeding the strongest reported conventional baseline, AV-MC, by 3.31 percentage points. These results support the effectiveness of HEAR for both sentiment polarity classification and sentiment intensity estimation.

Comparison with LLM-based Methods. Under the same LLM backbones, HEAR consistently outperforms MSE-Adapter across all five reported metrics. With Qwen-1.8B, HEAR improves Acc-2 from 80.44% to 81.55% and Acc2-w from 73.09% to 74.41%, while reducing MAE from 0.311 to 0.292. With LLaMA2-7B, Acc-2 increases from 75.53% to 77.74%, Acc2-w increases from 68.61% to 71.18%, and Corr improves from 0.553 to 0.619. With ChatGLM3-6B, HEAR improves Acc-2/F1 from 83.77%/83.76% to 85.20%/85.21%, yielding gains of 1.43 and 1.45 percentage points, respectively. It also improves Acc2-w from 75.24% to 77.85%, reduces MAE from 0.296 to 0.261, and increases Corr from 0.720 to 0.759. The improvements across all three backbones support the effectiveness of HEAR's structured evidence modeling and hierarchical reasoning.

Overall Analysis. Overall, HEAR-ChatGLM3-6B achieves the best reported results across all five metrics on CH-SIMS-V2: 85.20% Acc-2, 85.21% F1, 77.85% Acc2-w, 0.261 MAE, and 0.759 Corr. Its performance on Acc2-w indicates that the benefits extend to weak sentiment expressions, while the regression results support its effectiveness in estimating sentiment intensity. Together with the improvements over MSE-Adapter across three LLM backbones, these results complement the findings on CMU-MOSEI and support the applicability of hierarchical evidence reasoning to Chinese multimodal sentiment analysis.

Table 6: Comparison of sentiment prediction results on CH-SIMS-V2. The best and second-best results are highlighted in bold and underlined, respectively. * denotes results quoted from prior work.

Model	Acc-2↑	F1↑	Acc2-w↑	MAE↓	Corr↑
TFN* (Zadeh et al., 2017)	76.51	76.31	66.27	0.323	0.667
LMF* (Liu et al., 2018)	77.05	77.02	69.34	0.343	0.638
MulT* (Tsai et al., 2019)	79.50	79.59	69.61	0.317	0.703
MAG-BERT* (Rahman et al., 2020)	79.79	79.78	71.87	0.334	0.691
MISA (Hazarika et al., 2020)	80.53	80.63	70.50	0.314	0.725
Self-MM* (Yu et al., 2021)	79.01	78.89	71.87	0.335	0.640
MMIM (Han et al., 2021)	80.95	80.97	72.28	0.316	0.707
AV-MC (Liu et al., 2022)	82.50	82.55	74.54	0.297	0.732
CMamba (Li et al., 2024b)	83.40	83.50	–	<u>0.287</u>	<u>0.758</u>
MSE-Qwen-1.8B (Yang et al., 2025b)	80.44	80.24	73.09	0.311	0.678
MSE-LLaMA2-7B (Yang et al., 2025b)	75.53	75.44	68.61	0.382	0.553
MSE-ChatGLM3-6B (Yang et al., 2025b)	<u>83.77</u>	<u>83.76</u>	<u>75.24</u>	0.296	0.720
HEAR-Qwen-1.8B	81.55	81.46	74.41	0.292	0.701
HEAR-LLaMA2-7B	77.74	77.36	71.18	0.326	0.619
HEAR-ChatGLM3-6B	**85.20**	**85.21**	**77.85**	**0.261**	**0.759**

B.3 Extended Analysis for MELD and CHERMA

Results on MELD and CHERMA. On MELD, HEAR remains competitive with recent emotion recognition methods and achieves the best overall performance with ChatGLM3-6B, reaching 67.23% Acc and 66.31% WF1. Compared with the strongest existing baselines, it improves Acc over DKRF from 66.70% to 67.23% and WF1 over UniMSE from 65.51% to 66.31%. Similar improvements are observed over MSE-Adapter under the same LLM backbones. In particular, HEAR improves Acc/WF1 from 62.18%/59.87% to 63.90%/62.04% with Qwen-1.8B, and from 66.23%/65.13% to 67.23%/66.31% with ChatGLM3-6B.

On CHERMA, HEAR-ChatGLM3-6B achieves the best Acc and WF1 of 74.63% and 74.66%, respectively, outperforming MSE-ChatGLM3-6B by 1.73 and 1.93 percentage points. Notably, HEAR-Qwen-1.8B also achieves the second-best results despite its smaller backbone, with 73.82% Acc and 73.85% WF1. Across all three backbones, HEAR consistently improves over the corresponding MSE-Adapter variants, demonstrating that its effectiveness extends beyond sentiment intensity prediction to multi-class emotion recognition.

Overall Analysis. Together with the results on CMU-MOSEI and CH-SIMS-V2, the consistent improvements on MELD and CHERMA demonstrate the applicability of HEAR across different task formulations and languages. In particular, its gains on emotion recognition suggest that structured

evidence reasoning can effectively organize fine-grained multimodal cues for distinguishing discrete emotion categories, complementing its effectiveness in sentiment polarity and intensity prediction.

Table 7: Comparison of emotion recognition results on MELD and CHERMA. Acc and WF1 denote accuracy and weighted F1-score, respectively. The best and second-best results are highlighted in bold and underlined, respectively. * denotes results quoted from prior work.

Model	MELD		CHERMA	
	Acc↑	WF1↑	Acc↑	WF1↑
TFN* (Zadeh et al., 2017)	60.77	57.74	–	68.37
LMF* (Liu et al., 2018)	61.15	58.30	–	68.23
MulT (Tsai et al., 2019)	–	–	–	69.24
PMR (Lv et al., 2021)	–	–	–	69.53
LFMIM (Sun et al., 2023)	–	–	–	70.54
MMGCN* (Hu et al., 2021)	60.42	58.31	–	–
MM-DFN* (Hu et al., 2022a)	62.49	59.46	–	–
ConFEDE (Yang et al., 2023)	<u>63.80</u>	60.50	–	–
ALMT (Zhang et al., 2023)	<u>64.00</u>	61.00	–	–
SPAE (Yu et al., 2023a)	<u>63.90</u>	60.70	–	–
EmoCaps (Li et al., 2022)	–	64.00	–	–
GA2MIF (Li et al., 2024a)	61.65	58.94	–	–
UniMSE (Hu et al., 2022b)	65.09	<u>65.51</u>	–	–
UniSA$_{GPT2}$ (Li et al., 2023c)	48.12	31.26	–	–
UniSA$_{T5}$ (Li et al., 2023c)	64.52	62.17	–	–
UniSA$_{BART}$ (Li et al., 2023c)	62.34	62.22	–	–
DRKF (Jiang et al., 2025a)	<u>66.70</u>	65.40	–	–
MSE-Qwen-1.8B (Yang et al., 2025b)	62.18	59.87	70.38	70.21
MSE-LLaMA2-7B (Yang et al., 2025b)	65.14	63.66	71.58	71.41
MSE-ChatGLM3-6B (Yang et al., 2025b)	66.23	65.13	72.90	72.73
HEAR-Qwen-1.8B	63.90	62.04	<u>73.82</u>	<u>73.85</u>
HEAR-LLaMA2-7B	66.59	64.91	72.88	72.91
HEAR-ChatGLM3-6B	**67.23**	**66.31**	**74.63**	**74.66**

B.4 Additional Ablation Studies

Loss Function Ablation.

Table 8 evaluates the contribution of different training objectives. The full objective achieves the best performance across all reported metrics on both datasets. Using only the LLM prediction loss underperforms Full HEAR, supporting the benefit of the combined auxiliary supervision. On CH-SIMS-V2, removing $\mathcal{L}_{\text{prog}}$ causes the largest reductions in Acc-2 and Acc2-w, from 85.20% and 77.85% to 83.08% and 74.82%, respectively, corresponding to drops of 2.12 and 3.03 percentage points. Removing $\mathcal{L}_{\text{hier}}$ yields the lowest F1 of 82.94% and the highest MAE of 0.268, compared with 85.21% and 0.261 for Full HEAR. On MELD, removing $\mathcal{L}_{\text{hier}}$ produces the largest decreases in both Acc and WF1, from 67.23%/66.31% to 65.72%/65.07%. Removing $\mathcal{L}_{\text{disc}}$ also degrades all reported metrics on both datasets. These results support the contribution of each auxiliary objective within the full training framework.

Projection Strategy Ablation.

Figure 7 compares three strategies for projecting hierarchical evidence into the LLM. Direct concatenation yields the weakest performance across the reported metrics, while independent projection improves Acc-2 from 83.17% to 83.85% on CH-SIMS-V2 and Acc from 66.13% to 67.13% on MELD. Incremental projection further achieves the best results, reaching 85.20% Acc-2, 85.21% F1, 77.85% Acc2-w, 0.261 MAE, and 0.759 Corr on CH-SIMS-V2, together with 67.23% Acc and 66.31% WF1 on MELD. Compared with independent projection, it improves CH-SIMS-V2 Acc-2 by 1.35 percentage points and reduces MAE from 0.265 to 0.261. These results support progressively integrating evidence across reasoning levels when constructing the LLM inputs.

Table 8: Ablation study of different loss components on CH-SIMS-V2 and MELD. The best results are highlighted in bold. ↑ and ↓ indicate that higher and lower values are better, respectively.

Variant	CH-SIMS-V2					MELD	
	Acc-2↑	F1↑	Acc2-w↑	MAE↓	Corr↑	Acc↑	WF1↑
LLM Loss Only	83.87	83.86	76.44	0.265	0.755	66.86	66.07
w/o $\mathcal{L}_{\text{disc}}$	83.81	83.75	76.23	0.264	0.750	66.02	65.40
w/o $\mathcal{L}_{\text{hier}}$	83.09	82.94	75.19	0.268	0.756	65.72	65.07
w/o $\mathcal{L}_{\text{prog}}$	83.08	82.96	74.82	0.264	0.752	66.91	66.30
Full HEAR	**85.20**	**85.21**	**77.85**	**0.261**	**0.759**	**67.23**	**66.31**

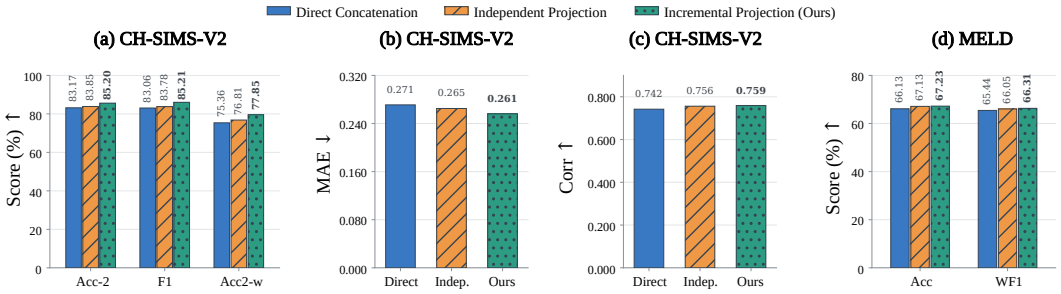


Figure 7: Comparison of different projection strategies on CH-SIMS-V2 and MELD. (a) Classification performance on CH-SIMS-V2; (b) MAE on CH-SIMS-V2; (c) Correlation on CH-SIMS-V2; and (d) classification performance on MELD. Incremental projection consistently achieves the best performance across all metrics.

B.5 Additional Hyperparameter Sensitivity

We further investigate the sensitivity of HEAR to four architectural hyperparameters, including the number of frequency bins N_f, the number of evidence tokens per view K, the hierarchical query counts (N_L, N_R, N_G), and the pseudo-token counts (M_L, M_R, M_G). As shown in Table 9, the default configuration consistently achieves the best performance across all reported metrics on the four benchmarks.

For N_f, increasing the number of frequency bins from 16 to 32 improves all reported metrics across the four datasets, whereas further increasing it to 64 provides no additional benefit. A similar pattern is observed for K, with four evidence tokens per view achieving the best results among the evaluated settings. On CH-SIMS-V2, reducing K from 4 to 2 decreases Acc-2 from 85.20% to 81.43%, a drop of 3.77 percentage points, and increases MAE from 0.261 to 0.268. Increasing K to 8 also underperforms the default setting, yielding 84.13% Acc-2 and 0.264 MAE. These results favor moderate spectral resolution and evidence token counts within the evaluated ranges.

The hierarchical query counts and pseudo-token counts exhibit a similar pattern. The intermediate query configuration $(16, 8, 3)$ consistently outperforms both the smaller $(8, 4, 2)$ and larger $(24, 12, 4)$ configurations, indicating that insufficient query capacity may limit hierarchical evidence reasoning, while simply increasing the number of queries does not necessarily improve performance. Likewise, $(M_L, M_R, M_G) = (4, 3, 2)$ provides the best overall results compared with both fewer and more pseudo-tokens. This suggests that a compact but sufficiently expressive set of pseudo-tokens is adequate for conveying hierarchical evidence to the frozen LLM.

Overall, the default configuration achieves the best reported results across all four benchmarks within the evaluated hyperparameter ranges. Increasing the numbers of frequency bins, evidence tokens, queries, or pseudo-tokens beyond the default settings does not improve performance. Smaller configurations also underperform the defaults, with a particularly pronounced reduction in CH-

Table 9: Hyperparameter sensitivity analysis on CH-SIMS-V2, CMU-MOSEI, MELD, and CHERMA. Each block varies one hyperparameter group while keeping the others fixed. † denotes the default configuration, whose results are shared across all blocks. The best result for each metric within each block is highlighted in bold. ↑ and ↓ indicate that higher and lower values are better, respectively.

Hyperparameter	Setting	CH-SIMS-V2		CMU-MOSEI		MELD		CHERMA	
		Acc-2↑	MAE↓	Acc-2↑	MAE↓	Acc↑	WF1↑	Acc↑	WF1↑
Frequency bins N_f	16	83.79	0.269	86.93	0.488	66.28	65.98	74.50	74.44
	$32^{\dagger}$	**85.20**	**0.261**	**87.29**	**0.487**	**67.23**	**66.31**	**74.63**	**74.66**
	64	83.75	0.267	87.21	0.489	67.01	66.05	74.58	74.53
Evidence tokens per view K	2	81.43	0.268	86.50	0.491	66.09	66.01	73.50	73.37
	$4^{\dagger}$	**85.20**	**0.261**	**87.29**	**0.487**	**67.23**	**66.31**	**74.63**	**74.66**
	8	84.13	0.264	87.24	0.501	66.97	65.95	74.37	74.38
Query counts (N_L, N_R, N_G)	$(8, 4, 2)$	83.95	0.266	87.14	0.506	66.82	65.67	73.57	73.83
	$(16, 8, 3)^{\dagger}$	**85.20**	**0.261**	**87.29**	**0.487**	**67.23**	**66.31**	**74.63**	**74.66**
	$(24, 12, 4)$	84.14	0.272	86.84	0.489	66.40	65.78	74.10	73.77
Pseudo-token counts (M_L, M_R, M_G)	$(2, 1, 1)$	83.27	0.265	86.95	0.491	67.19	66.28	74.47	74.38
	$(4, 3, 2)^{\dagger}$	**85.20**	**0.261**	**87.29**	**0.487**	**67.23**	**66.31**	**74.63**	**74.66**
	$(8, 4, 3)$	84.24	0.264	87.08	0.489	66.86	66.23	74.55	74.40

SIMS-V2 Acc-2 when using only two evidence tokens per view. These findings support the selected configuration as an effective choice among the tested settings.

C Additional Visualization Analysis

C.1 Visualization of Hierarchical Evidence Reasoning

To qualitatively examine the hierarchical reasoning process of HEAR, we visualize two representative examples from each of CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA. Figures 8–15 present the interaction patterns at the local, relational, and global reasoning stages, together with the corresponding textual input, ground-truth label, and model prediction.

Hierarchical Reasoning Patterns. Across different datasets and examples, the visualizations exhibit a consistent progression from fine-grained evidence selection to relational interaction and global abstraction. At the local stage, different queries respond differently to multimodal semantic evidence groups, indicating that HEAR preserves heterogeneous evidence rather than immediately collapsing multimodal information into a single representation. The relational stage further reorganizes the selected local evidence through inter-evidence interactions, producing structured response patterns across local evidence tokens. Finally, the global stage integrates information from the preceding reasoning levels into a compact set of global evidence representations. These progressively transformed patterns qualitatively illustrate the local-to-relational-to-global reasoning process of HEAR.

Cross-Dataset Observations. Despite differences in language, label space, and task formulation, similar hierarchical organization patterns can be observed across CMU-MOSEI, CH-SIMS-V2, MELD, and CHERMA. The regression examples from CMU-MOSEI and CH-SIMS-V2 show that the progressively abstracted evidence supports predictions close to the corresponding sentiment intensities, while the MELD and CHERMA examples illustrate the same reasoning process for discrete emotion recognition. Together with the quantitative ablation results, these visualizations provide complementary qualitative evidence that HEAR progressively organizes multimodal evidence across reasoning levels rather than directly aggregating all evidence into a single global representation.

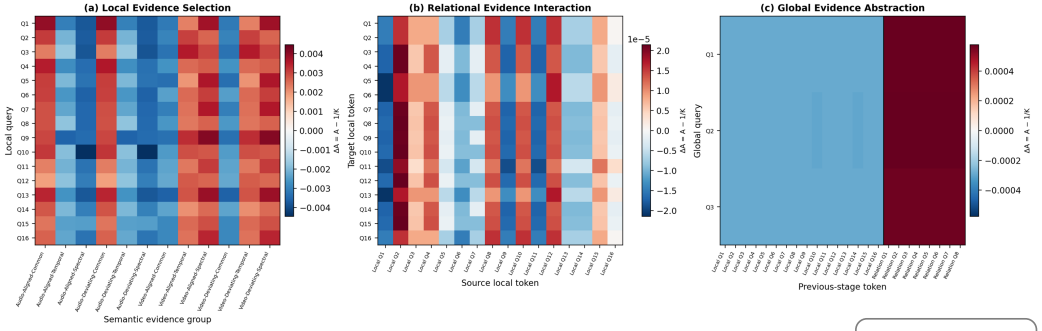


Figure 8: Visualization of hierarchical evidence reasoning in HEAR on a CMU-MOSEI example. (a) Local evidence selection over multimodal semantic evidence groups; (b) relational interactions among local evidence; and (c) global evidence abstraction over local and relational representations. The ground-truth sentiment intensity is $+0.333$, while HEAR predicts $+0.434$. Colors indicate stage-specific interaction strengths.

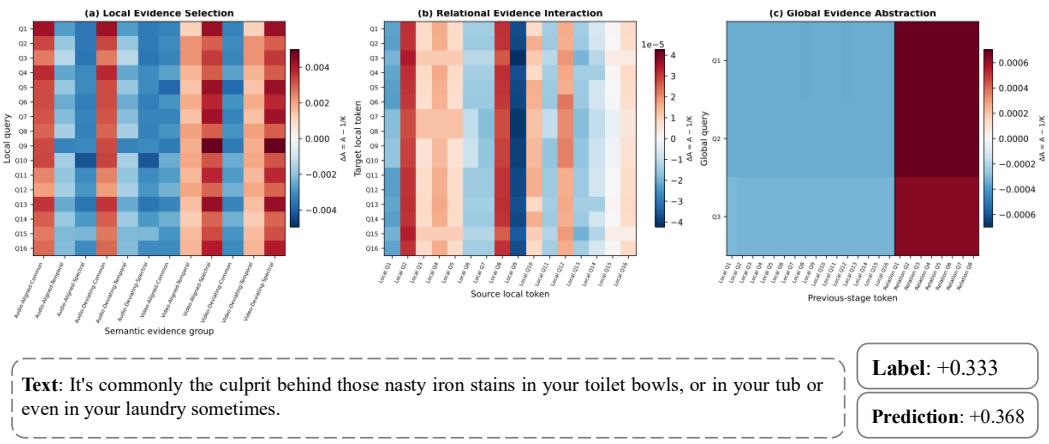


Figure 9: Visualization of hierarchical evidence reasoning in HEAR on a CMU-MOSEI example. (a) Local evidence selection over multimodal semantic evidence groups; (b) relational interactions among local evidence; and (c) global evidence abstraction over local and relational representations. The ground-truth sentiment intensity is $+0.333$, while HEAR predicts $+0.368$. Colors indicate stage-specific interaction strengths.

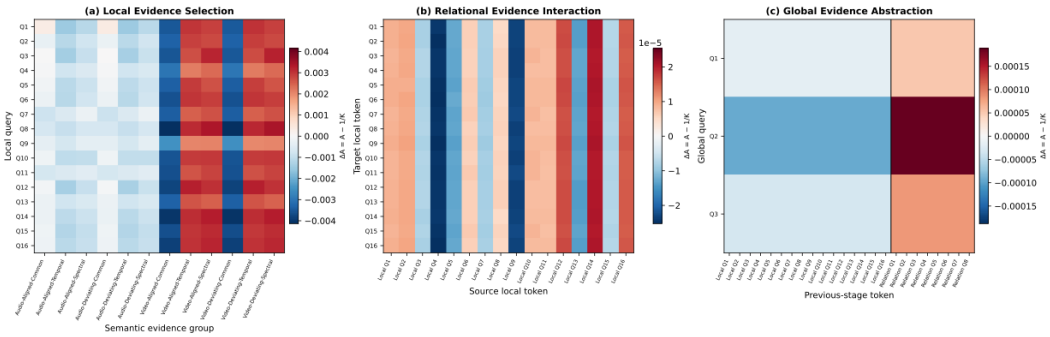


Figure 10: Visualization of hierarchical evidence reasoning in HEAR on a CH-SIMS-V2 example. (a) Local evidence selection over multimodal semantic evidence groups; (b) relational interactions among local evidence; and (c) global evidence abstraction over local and relational representations. The ground-truth sentiment intensity is $+1.0$, while HEAR predicts $+0.909$. Colors indicate stage-specific interaction strengths.

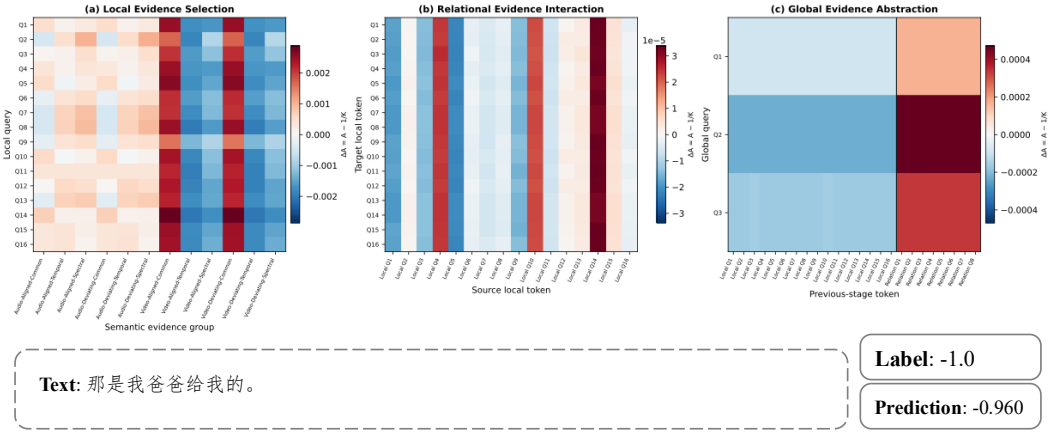
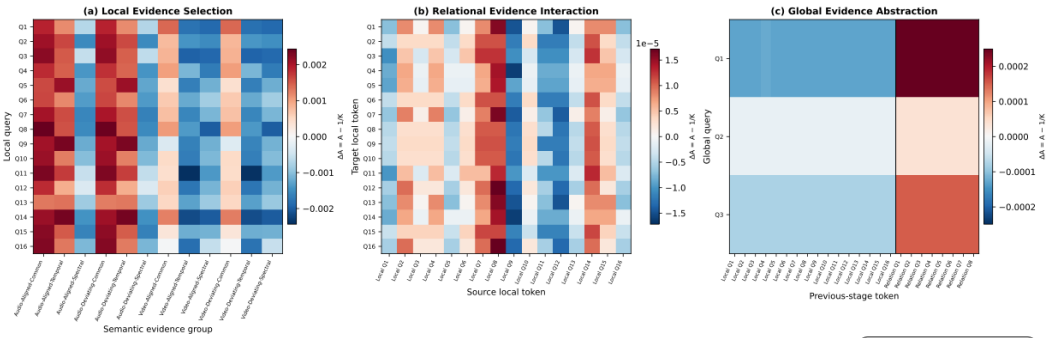


Figure 11: Visualization of hierarchical evidence reasoning in HEAR on a CH-SIMS-V2 example. (a) Local evidence selection over multimodal semantic evidence groups; (b) relational interactions among local evidence; and (c) global evidence abstraction over local and relational representations. The ground-truth sentiment intensity is -1.0, while HEAR predicts -0.960. Colors indicate stage-specific interaction strengths.

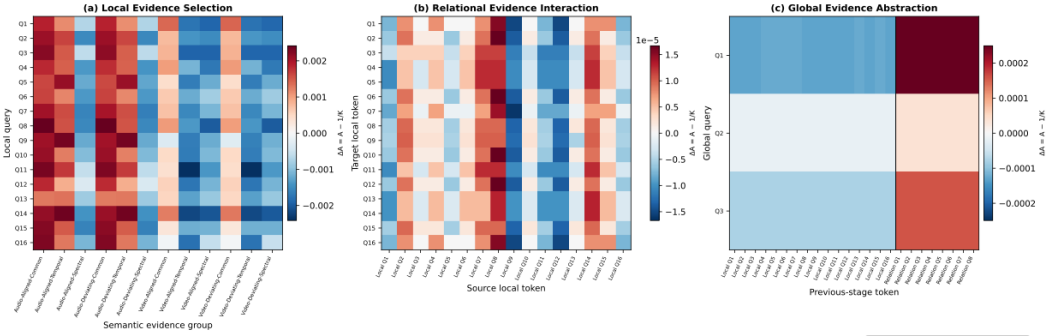


Text: Ross You can do it! Ross You can do it! Ross </s> Ross Okay. </s> Ross You got it?

Label: 0

Prediction: 0

Figure 12: Visualization of hierarchical evidence reasoning in HEAR on a MELD example. (a) Local evidence selection over multimodal semantic evidence groups; (b) relational interactions among local evidence; and (c) global evidence abstraction over local and relational representations. The ground-truth emotion label is 0, while HEAR predicts 0. Colors indicate stage-specific interaction strengths.

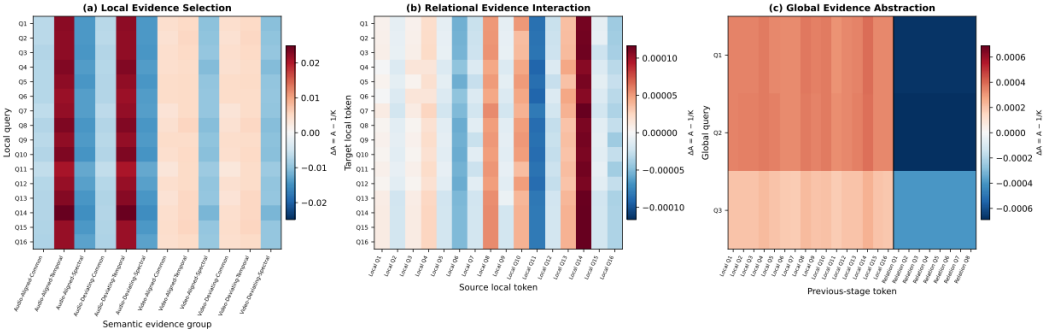


Text: Joey Let's lwas just yeah, right. joey Push! </s> joey Push!

Label: 4

Prediction: 4

Figure 13: Visualization of hierarchical evidence reasoning in HEAR on a MELD example. (a) Local evidence selection over multimodal semantic evidence groups; (b) relational interactions among local evidence; and (c) global evidence abstraction over local and relational representations. The ground-truth emotion label is 4, while HEAR predicts 4. Colors indicate stage-specific interaction strengths.

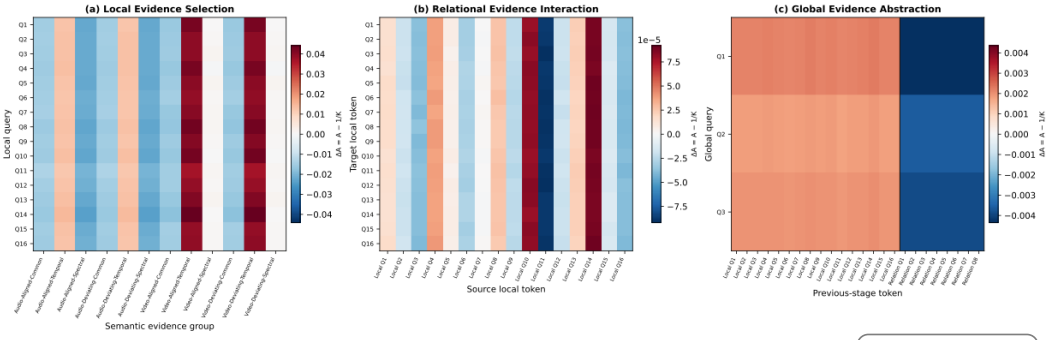


Text: 现在黎明朗和江百合在一起，但是我的孩子没了。

Label: 5

Prediction: 5

Figure 14: Visualization of hierarchical evidence reasoning in HEAR on a CHERMA example. (a) Local evidence selection over multimodal semantic evidence groups; (b) relational interactions among local evidence; and (c) global evidence abstraction over local and relational representations. The ground-truth emotion label is 5, while HEAR predicts 5. Colors indicate stage-specific interaction strengths.



Text: 学校四点钟放学，现在已经过去这么久了。

Label: 2

Prediction: 2

Figure 15: Visualization of hierarchical evidence reasoning in HEAR on a CHERMA example. (a) Local evidence selection over multimodal semantic evidence groups; (b) relational interactions among local evidence; and (c) global evidence abstraction over local and relational representations. The ground-truth emotion label is 2, while HEAR predicts 2. Colors indicate stage-specific interaction strengths.