Who Should I Trust? Explicit Confidence-Focused Multimodal Intent Recognition

Yi Liu, Qimeng Yang*, Lanlan Lu

Xinjiang University
Urumqi 830046, Xinjiang, China
15823785909@163.com, yangqm@xju.edu.cn, 107552304950@stu.xju.edu.cn

Abstract

Multimodal intent recognition is aimed at understanding user intentions by integrating information from multiple modalities. It has attracted increasing attention in recently developed dialog systems. The existing studies have focused mainly on modeling semantic interactions within and across modalities, but they often overlook the reliability of each modality. In real-world scenarios, inputs may be corrupted by noisy audio, blurred or occluded videos, or ambiguous text, making it difficult for the employed model to determine who to trust and how much to trust. To address this challenge, we propose a method called explicit confidence-focused multimodal intent recognition (ECFMIR). The core idea of this approach is to assign each modality and each cross-modal associations feature a dedicated confidence lens (CLens) that explicitly estimates the confidence level in a hypothetical manner. This design helps reduce the degree of uncertainty and mitigate the risk of incorrect predictions when addressing conflicting inputs. Comprehensive experiments conducted on two benchmark multimodal intent recognition datasets demonstrate the effectiveness of our method. A further analysis reveals that ECFMIR achieves significant advantages for high-conflict categories and under low-resource conditions.

Code — https://github.com/SixOne666111/ECFMIR

Introduction

Intent recognition in artificial intelligence is aimed at identifying the underlying purposes behind user inputs (Chen, Zhuo, and Wang 2019). Prior research has extensively demonstrated the importance of textual modalities (Tsai et al. 2018). However, beyond textual descriptions, visual cues and acoustic signals also carry rich information and often exhibit strong complementarity (Liu et al. 2022). The incorporation of such modalities is not only beneficial but also more practical in real-world applications (Zhao, Zhang, and Geng 2024). Multimodal intent recognition goes beyond traditional unimodal approaches by leveraging complementary data derived from text, speech, and vision (Sun et al. 2024). This holistic perspective enables models to gain a more comprehensive and nuanced understanding of user intents (Zhang et al. 2024b).

*Corresponding Author.

Figure 1: Modalities may conflict or exhibit varying degrees of reliability.

Despite the progress made, most existing work overlooks the reliability of each modality (Ma et al. 2023). Some studies focus solely on intra- or intermodal fusion, whereas others rely on complex attention mechanisms to implicitly handle noisy inputs (He et al. 2024). Few studies have provided an interpretable and explicit confidence estimate for each modality as a whole (Zhao and Li 2024). As a result, the various contributions of different modalities are often underappreciated.

Some modalities may exhibit low apparent noise levels but contain misleading information, making naive fusion insufficient for enabling a model to select reliable cues on the basis of appropriate confidence, thereby undermining the reliability of intent recognition. As illustrated in Figure 1, a middle-aged man is speaking with his brow furrowed tightly, clearly conveying suspicion and dissatisfaction. He utters the phrase "I guess this is just okay" in a flat tone, while his voice contrasts sharply it is marked by emphatic stress, an elevated pitch, and an accelerated pace. This scene presents a pronounced multimodal conflict: the facial and body language of the man signal negation and rejection, and his audio conveys intense emphasis, whereas the textual content appears weak and subdued. Under such conditions, models struggle to discern which modality is trustworthy and accurate.

To address the above challenges, we propose a novel framework, explicit confidence-focused multimodal intent recognition (ECFMIR), which involves constructing cross-modal associative features and explicitly modeling their reliability. At the core of ECFMIR is a confidence lens (CLens), which estimates the uncertainty of each feature and trans-

forms it into an unreliability index (UI). This UI is then modulated by a learnable focal length derived from the latent representations of the features and further converted into a confidence score through an exponentially decaying function. As a result, features with higher unreliability values receive lower confidence, enabling the model to make more robust decisions.

The resulting confidence scores directly guide a sample-level modality selection process, allowing the model to determine which modalities and cross-modal relations are trustworthy. Extensive experiments demonstrate that our method consistently outperforms the state-of-the-art baselines. Our main contributions are as follows.

- We introduce a CLens, which is a novel mechanism that estimates the UI for each modality and regulates it via focal length modulation, enabling explicit and interpretable modality reliability modeling.
- We propose a multimodal intent recognition framework that explicitly models the reliability of both individual modalities and their cross-modal associations. To our knowledge, this is the first work to provide such a reliability estimation procedure in the context of multimodal intent recognition.
- We conduct comprehensive experiments on two challenging benchmark datasets, which demonstrate that our method outperforms the existing state-of-the-art approaches in multimodal intent recognition tasks.

Related Work

Multimodal Intent Recognition

Multimodal intent recognition enhances semantic understanding by integrating text, visual, and audio signals. Early approaches such as MulT (Bhattacharjee et al. 2022), introduced directional cross-modal attention for performing unaligned sequence fusion. MISA (Hazarika, Zimmermann, and Poria 2020) disentangles modality-invariant and modality-specific features to reduce distribution gaps. The multimodal adaptation gate-bidirectional encoder representations from transformers (MAG-BERT) (Rahman et al. 2020) uses an MAG to inject audio–visual cues into pretrained language models, achieving near-human performance on CMU-MOSI.

Recent advances have addressed data scarcity and noisy modalities. InMu-Net (Zhu et al. 2024) employs an information bottleneck and regularizations to filter redundancy and provide increased robustness. SDIF-DA (Huang et al. 2024) combines shallow alignment with deep fusion and leverages ChatGPT to augment data, setting new benchmarks. TCL-MAP (Zhou et al. 2024) combines modality-aware prompt learning with token-level contrastive learning to represent intents in a refined manner. MVCL-DAF (Hu et al. 2025) further introduced dynamic attention fusion and multiview contrastive learning, demonstrating strong generalizability across datasets.

Trustworthy Multimodal Learning

Confidence modeling has emerged as the key to improving the robustness and interpretability of multimodal learning. UNO (Tian et al. 2020) unifies uncertainty modeling via deviation ratios and Noisy-Or fusion to achieve robust inference. MLCLNet (Zheng et al. 2023) was proposed as a three-level confidence framework for addressing noise suppression and cross-modal consistency. DPNET (Zou et al. 2023) systematically models confidence across different modalities, labels, and structures via adaptive weighting and low-rank tensor methods.

To perform unsupervised fusion, COLD Fusion (Tellamekala et al. 2023) models modality uncertainties with Gaussian distributions and learns adaptive fusion weights. TMSON (Xie et al. 2024) extends this process with Bayesian fusion and an ordinal-aware triplet loss to enforce ranking consistency and improve the ability to detect misclassifications.

Unlike previously developed approaches that rely on architectural tweaks or Bayesian inference, our method directly learns confidence distributions from feature variance, compresses unreliability via an exponential index, and combines entropy with a learnable focal coefficient. This enables confidence estimation to be effectively performed without additional networks or complex inferences, significantly reducing the level of overconfidence.

Method

Previous studies mainly focused on fusion strategies driven by cross-modal interactions, while often neglecting the varying reliability of different modalities. Inspired by trustworthy multimodal learning, we propose ECFMIR, as illustrated in Figure 2. The key idea is that, for each modality, it is crucial not only to understand what it conveys but also how much it can be trusted; and for cross-modal relations, to determine not only whether they exist but also whether they are complementary or conflicting. From an information-theoretic perspective, ECFMIR estimates the discriminative power of each modality, achieving more robust and reliable multimodal fusion.

Method Pipeline

Feature Extraction Given a set of input samples comprising a text sequence T, an audio signal A, and a video frame sequence V, we first extract context-aware semantic representations $X^{(T)}$ from text using a pretrained BERT model (Devlin et al. 2019), high-dimensional acoustic features $X^{(A)}$ from raw audio via Wav2vec2.0 (Baevski et al. 2020), and frame-level spatiotemporal visual features $X^{(V)}$ using a Swin-Transformer (Liu et al. 2021):

$$T \xrightarrow{\text{BERT}} X^{(T)} \in \mathbb{R}^{L_T \times d_T}, \tag{1}$$

$$A \xrightarrow{\text{Wav2vec2.0}} X^{(A)} \in \mathbb{R}^{L_A \times d_A}, \tag{2}$$

$$V \xrightarrow{\text{Swin-Transformer}} X^{(V)} \in \mathbb{R}^{L_V \times d_V}. \tag{3}$$

The respective sequence lengths and feature dimensions for the three modalities are denoted by L_T, L_A, L_V and d_T, d_A, d_V.

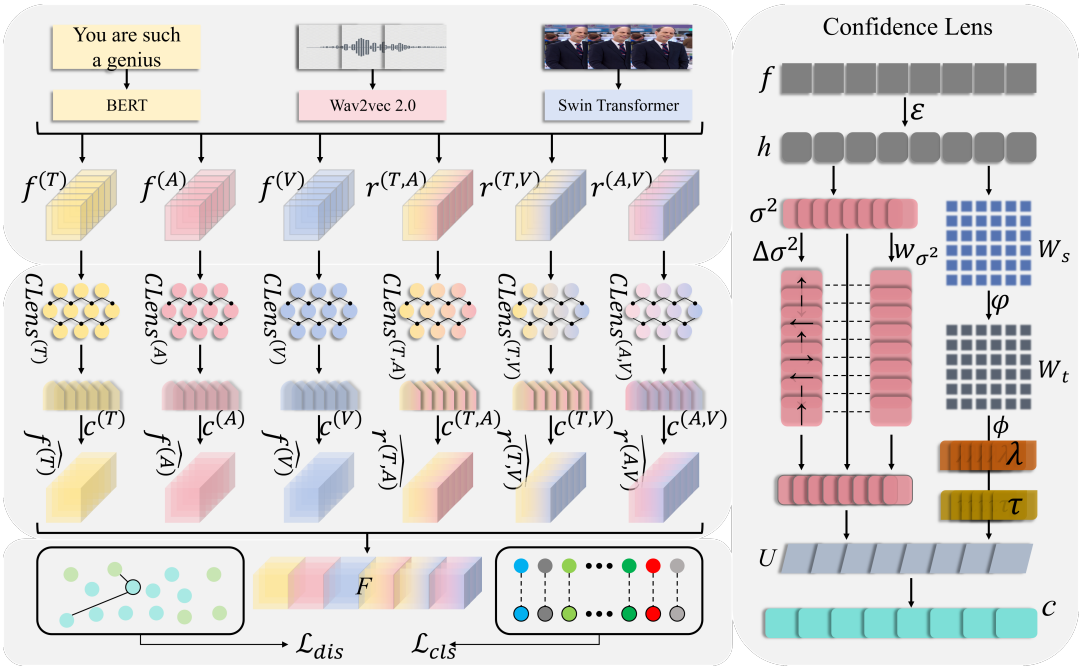

Figure 2: Overview of the ECFMIR architecture. Given a multimodal input, modality-specific and cross-modal features are first constructed. Each feature is then evaluated by a dedicated CLens to estimate its reliability, which guides the confidence-aware fusion process to perform the final intent recognition step.

The features of each modality are then projected into a shared d-dimensional semantic space through modality-specific linear projection matrices $W^{(T)}, W^{(A)}, W^{(V)}$. Applying mean pooling yields the following compact unimodal representations:

$$f^{(m)} = \text{Mean}\big(W^{(m)} X^{(m)}\big), \quad m \in \{T, A, V\} \tag{4}$$

where $W^{(T)} \in \mathbb{R}^{d \times d_T}$, $W^{(A)} \in \mathbb{R}^{d \times d_A}$, and $W^{(V)} \in \mathbb{R}^{d \times d_V}$ are the projection matrices. To capture cross-modal associations, we construct cross-modal associations feature vectors by concatenating the corresponding unimodal representations:

$$r^{(m_i, m_j)} = \left[f^{(m_i)}; f^{(m_j)}\right] \tag{5}$$

where (m_i, m_j) ranges over the modality pairs $\{(T, A), (T, V), (A, V)\}$.

Confidence Perception and Fusion We design the CLens module (detailed below) to address the two core questions raised at the outset of this work.

To quantify how much we should trust a modality, we apply CLens units to both unimodal representations and cross-modal associations features. Specifically, for each modality $m \in \{T, A, V\}$ and cross-modal associations features $m_i, m_j) \in \{(T, A), (T, V), (A, V)\}$, the corresponding confidence scores are computed as follows:

$$c^{(m)} = \text{CLens}^{(m)}(f^{(m)}) \tag{6}$$

$$c^{(m_i, m_j)} = \text{CLens}^{(m_i, m_j)}(r^{(m_i, m_j)}) \tag{7}$$

These confidence scores explicitly measure the reliability of each feature, enabling sample-level confidence modulation. Accordingly, the weighted reliable feature representations for the unimodal and cross-modal features are as follows:

$$\hat{f}^{(m)} = c^{(m)} \cdot f^{(m)} \tag{8}$$

$$\hat{r}^{(m_i, m_j)} = c^{(m_i, m_j)} \cdot r^{(m_i, m_j)} \tag{9}$$

To maximize the degree of information retention, we concatenate all confidence-weighted features to form a fused representation:

$$\mathcal{F} = \text{Concat}\big(\{\hat{f}^{(m)}\}, \{\hat{r}^{(m_i, m_j)}\}\big) \in \mathbb{R}^{9d} \tag{10}$$

Decision Making To encourage the model to learn a more discriminative feature space, we introduce a top-k distance loss. Specifically, for each sample contained in the representation set F, we select the k_p most distant positive samples (i.e., those with the same label) and the k_n nearest negative samples (i.e., those with different labels). The distance-based loss is defined as shown below:

$$\mathcal{L}_{\text{dis}} = \frac{1}{B k_p k_n} \sum_{i=1}^{B} \sum_{j=1}^{k_p} \sum_{l=1}^{k_n} \ln\left(1 + d(f_i, f_j) - d(f_i, f_l)\right) \tag{11}$$

where B denotes the batch size and $d(\cdot, \cdot)$ is the Euclidean distance metric.

The final fused representation F is fed into a multilayer perceptron (MLP) followed by a Softmax function to produce the predicted intent distribution:

$$y = \text{Softmax}(\text{MLP}(F)) \in \mathbb{R}^{C} \tag{12}$$

where C is the number of intent categories.

To optimize the classification performance of the model, we employ the cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\log \hat{y}_{i,y_i} \tag{13}$$

where $\hat{y}_{i,y_i}$ is the predicted probability of the i-th sample belonging to the ground-truth class y_i.

The overall training objective combines classification and distance losses:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dis}} \tag{14}$$

Confidence Lens

The CLens module is not designed to extract semantic content or denoise input features. Instead, its primary objective is to assess the reliability of any given input feature vector f.

Representation Focusing Given an input feature vector $f \in \mathbb{R}^D$, we first encode it using an MLP equipped with residual connections, projecting f into a latent space to obtain an abstract representation h for reliability estimation purposes:

$$h = \varepsilon(f) \in \mathbb{R}^D \tag{15}$$

The uncertainty of each latent dimension in h is captured by a variance vector σ^2, where each element reflects the confidence fluctuation associated with the corresponding dimension:

$$\sigma^2 = \text{Softplus}(\text{MLP}_\sigma(h)) \tag{16}$$

Here, Softplus activation ensures the nonnegativity of the estimated variances, enabling a stable and interpretable confidence modeling process.

Geometric Calibration Uncertainty estimation and confidence modeling are often affected by overconfidence, with initial variance predictions typically underestimated—especially for challenging samples. Geometric correction addresses this issue by enabling the model to identify underestimated uncertainty patterns within the variance vector. By leveraging a sigmoid gating function, it adaptively scales the original variance to produce more reliable uncertainty estimates.

Our correction process does not directly scale σ^2; instead, it decomposes the update into two independent components: a direction and a magnitude. This decoupling strategy affords greater flexibility and finer control. Specifically, the model first determines an optimal update direction on the basis of the current latent parameter state. We define an update vector $\Delta\sigma^2$, which is solely a function of the current variance σ^2:

$$\Delta\sigma^2 = \text{P}_\Delta(\sigma^2) \tag{17}$$

Here, $\mathbf{P}_\Delta$ denotes a linear transformation matrix that is designed to learn the geometric directions of specific bias signals. However, not all dimensions require uniform adjustments; some of them may already be optimally positioned, whereas others may demand substantial corrections. To address this situation, we introduce a focal scaling factor $\mathbf{w}_{\sigma^2}$, which acts as an independent gating mechanism for each dimension of the update vector:

$$\mathbf{w}_{\sigma^2} = \phi\big(\text{P}_w(\sigma^2)\big) \in (0,1) \tag{18}$$

Similarly, $\mathbf{P}_w$ denotes another linear transformation matrix, and ϕ represents the sigmoid function. Each element of the gating vector $\mathbf{w}_{\sigma^2}$ lies within the interval $(0,1)$, indicating its modulation strength rather than its direction. Values close to 1 permit substantial adjustments, whereas those near 0 effectively preserve the original state. This mechanism imparts essential locality and sparsity characteristics to the refinement process. Finally, we combine the update vector with the gating signal to produce the corrected variance vector:

$$\hat{\sigma}^2 = \sigma^2 + \mathbf{w}_{\sigma^2} \odot \Delta\sigma^2 \tag{19}$$

where $\odot$ denotes the elementwise multiplication operation. To prevent numerical instability, a lower bound constraint is applied:

$$\hat{\sigma}^2 \leftarrow \max(\hat{\sigma}^2, \epsilon) \tag{20}$$

where ϵ is a small positive constant.

Utilizing the corrected latent variance $\hat{\sigma}^2$, we define a scalar UI U for the hidden representation h by aggregating the uncertainty across all latent dimensions and smoothing its effect via a logarithm to prevent extreme values from dominating:

$$U = \log\left(1 + \frac{1}{d_z}\sum_{i=1}^{d_z}\hat{\sigma}_i^2\right) \tag{21}$$

Focal Modulation Different samples may exhibit varying sensitivities to unreliability. For example, a minor uncertainty in a simple sample might be negligible, whereas an uncertainty in a complex sample with the same magnitude could signal a potential risk. To address this issue, we predict a focal coefficient T on the basis of the hidden representation $\mathbf{h}$, which modulates the influence of the UI. This coefficient dynamically adjusts itself according to the complexity encoded in $\mathbf{h}$:

$$T = \tau + \lambda \cdot \phi\big(\mathbf{W}_t \cdot \phi(\mathbf{W}_s \cdot \mathbf{h})\big) \tag{22}$$

Here, ϕ denotes the GELU activation function, τ is the lower bound of the focal coefficient, λ controls its adjustment range, and $\mathbf{W}_t$, $\mathbf{W}_s$ are learnable projection matrices. This focal coefficient enables the model to dynamically adapt its sensitivity to unreliability based on the characteristics of the given samples, thereby achieving more flexible confidence estimation.

Derivation of Confidence Scores Ultimately, we transform the unreliability index into a normalized confidence score using a focal modulation exponential decay function. This score directly addresses the question of "to what extent should we trust this information":

$$c = \exp(-T \cdot U), \quad c \in (0,1] \tag{23}$$

Methods	MIntRec				MIntRec2.0			
	ACC (%)	WF1 (%)	WP (%)	WR (%)	ACC (%)	F1 (%)	P (%)	R (%)
MulT	71.46	71.10	71.65	71.46	57.75	51.34	52.73	50.83
MAG-BERT	69.89	69.26	70.09	69.89	55.98	50.29	51.49	50.16
MISA	71.46	72.06	73.26	71.46	56.76	51.76	52.54	51.58
MVCL-DAF	71.01	71.02	71.74	71.01	56.47	50.96	51.57	50.94
SDIF	71.46	71.43	71.66	71.46	57.65	52.16	53.70	52.29
TCL-MAP	72.36	72.21	73.11	72.36	45.11	44.69	45.70	45.11
ECFMIR	**74.38**	**74.51**	**75.05**	**74.38**	**58.58**	**53.39**	**54.22**	**54.39**

Table 1: Performance comparison among different methods on the MIntRec and MIntRec2.0 datasets.

This formulation integrates the unreliability index with a sample-specific focal parameter. When the unreliability index is high, the confidence score c correspondingly decreases, prompting the model to adopt a cautious stance toward highly uncertain samples and thus avoiding overconfidence.

Experiments

Datasets and Metrics

We evaluate our method on two challenging intent recognition datasets. MIntRec (Zhang et al. 2022) contains 2,224 samples with 20 intent labels across text, visual, and acoustic modalities. MIntRec2.0 (Zhang et al. 2024a), a large-scale benchmark for multimodal intent recognition, includes 1,245 dialogues and 15,040 samples integrating text, video, and audio. It further incorporates naturally occurring out-of-context samples, offering a more rigorous evaluation.

To comprehensively evaluate model performance, we adopt ACC, WF1, WP, and WR on MIntRec, where weighted metrics mitigate class imbalance, and ACC, F1, P, and R on MIntRec2.0.

Experimental Settings

Our model employs a learning rate of 2.5×10^{-5}. The hyperparameters are set as $\tau = 0.6$ and $\lambda = 2$. For a fair comparison, all methods utilize bert-base-uncased (Devlin et al. 2019) and wav2vec2-base-960h (Baevski et al. 2020) to extract textual and audio features, respectively, whereas the Swin-Transformer (Liu et al. 2021) is used for visual feature extraction. Optimization is performed using the AdamW optimizer (Loshchilov and Hutter 2017). The training batch size is set to 16, with the validation and testing batch sizes fixed to 8. All the experiments are conducted on an NVIDIA GeForce RTX 4090 GPU.

Baselines

In our experiments, we compare our method with six state-of-the-art baselines: (1) MulT (Bhattacharjee et al. 2022), which was developed for handling unaligned multimodal sequences via cross-modal attention; (2) MAG-BERT (Rahman et al. 2020), which integrates an MAG into BERT without modifying its backbone; (3) MISA (Hazarika, Zimmermann, and Poria 2020), which jointly learns modality-invariant and modality-specific subspaces to balance shared and unique features; (4) TCL-MAP (Zhou et al. 2024), which enhances textual features via modality-aware prompting and cross-modal alignment; (5) SDIF (Huang et al. 2024), which employs a shallow-to-deep interaction strategy for achieving progressive modality integration; and (6) MVCL-DAF (Hu et al. 2025), which uses dynamic attention fusion and contrastive learning to adaptively weight modal features.

Main Results

The experimental results are summarized in Table 1, where the best scores obtained in terms of each metric are highlighted in bold. Our method outperforms all existing state-of-the-art approaches across all the evaluation metrics. On the MIntRec dataset, our approach achieves an ACC of 74.38%, a WF1 of 74.51%, a WP of 75.05%, and a WR of 74.38%, surpassing the metrics of the baseline models by 2.02%, 2.3%, 1.79%, and 2.02%, respectively. Despite the MIntRec 2.0 dataset containing up to 38.28% out-of-context samples with unknown labels, our method still demonstrates substantial improvements on this dataset, attaining 58.58% ACC, 53.39% F1, 54.22% P, and 54.39% R. These results represent gains of 0.83%, 1.23%, 0.52%, and 2.1% over the strongest baselines, respectively.

Ablation Experiments

Table 2 shows the ablation results obtained with respect to the key components of ECFMIR. Six configurations are evaluated to assess the contribution of each module.

In (a), the unimodal confidence weights are removed; in (b), the pairwise weights are excluded. Both variants lead to clear performance drops, especially on MIntRec2.0 (the F1 values decrease to 50.24% and 50.50%), highlighting the necessity of jointly modeling unimodal and cross-modal confidence.

Configurations (c) and (d) test two CLens strategies. Removing the focal adjustment scheme in (c) impairs the resulting performance (F1 $\downarrow$ 51.38%, R $\downarrow$ 51.07%), showing its effectiveness in terms of handling overconfident or uncertain regions. Similarly, removing the geometric focusing approach in (d) degrades the results, confirming its role in capturing variance-aware reliability.

In (e), excluding the top-k distance loss slightly reduces the resulting accuracy, indicating its utility for producing a better structure and achieving superior generalization.

Setting	MIntRec				MIntRec2.0			
	ACC (%)	WF1 (%)	WP (%)	WR (%)	ACC (%)	F1 (%)	P (%)	R (%)
(a) Only Intermodal Confidence	70.34	70.69	72.98	70.34	55.68	50.24	51.62	49.91
(b) Only Unimodal Confidence	71.91	72.12	73.03	71.91	56.81	50.5	51.9	50.1
(c) w/o Focal Modulation	71.91	72.14	73.18	71.91	57.11	51.38	52.71	51.07
(d) w/o Geometric Calibration	72.36	72.73	73.83	72.36	58.19	52.57	53.76	51.99
(e) w/o top-k distance Loss $\mathcal{L}_{dis}$	73.03	72.83	73.11	73.03	57.3	50.86	51.92	50.87
(f) ECFMIR	**74.38**	**74.51**	**75.05**	**74.38**	**58.58**	**53.39**	**54.22**	**54.39**

Table 2: Performance comparison results obtained under different settings on the MIntRec and MIntRec2.0 datasets.

The full model (f) outperforms all the variants, validating the robustness and effectiveness of our method.

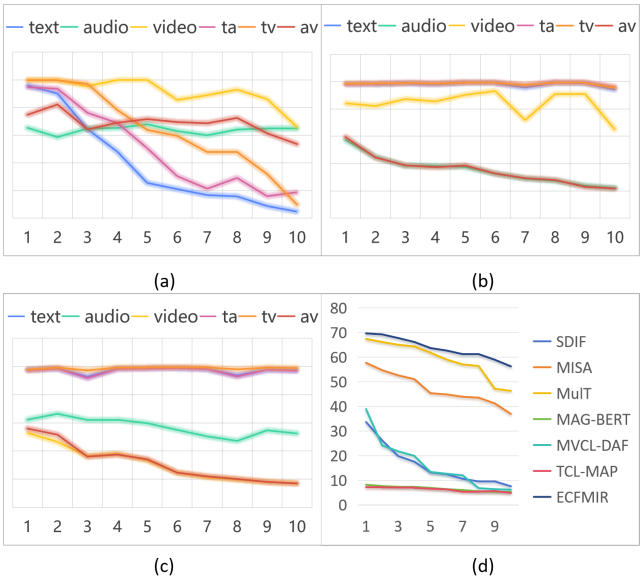

Figure 3: Confidence behavior exhibited under the injection of modality-specific noise on the MintRec dataset: (a) noise added to text, (b) audio, and (c) video features; (d) the performance achieved by the model under the injection of text noise, demonstrating its robustness in low-resource conditions.

Confidence Calibration Experiment

To evaluate the quality of the confidence estimates produced by our model, we employ three widely used calibration error metrics: the expected calibration error (ECE), average calibration error (ACE), and confidence error (OE). These metrics quantify the discrepancy between the predicted confidence and the actual accuracy, where lower values indicate better calibration effects and more reliable confidence estimates. As shown in Table 4, ECFMIR outperforms all the compared methods across these metrics, achieving a 15.99% ECE, a 15.82% ACE, and a 20.92% OE–the lowest values overall–demonstrating its superior confidence prediction accuracy and its ability to effectively mitigate overconfidence.

Noise Analysis

To assess the ability of our model to estimate the reliability of different modalities, we conduct noise injection experiments on the MintRec dataset. Gaussian noise with zero mean and variance U is added to each modality. For text, an additional variance of 10 is used to amplify the perturbation effect because of the typically higher density and stability of this modality. Figures 3(a)–(c) show that increasing the noise reduces the confidence in the affected modality, prompting the model to shift its focus to more reliable sources and leverage cross-modal cues. Figure 3(d) illustrates the robustness exhibited by the model under heavy text noise, highlighting its capacity to adaptively prioritize trustworthy modalities in noisy or low-resource settings.

Modal Conflict Analysis

To evaluate the performance of the model on intent categories with modality conflicts, we compute the cosine similarity between the fused prediction and each unimodal prediction for every sample in the training set and then aggregate these similarities by category to obtain an average conflict score per class. All categories are subsequently ranked by their conflict scores, and the ten categories with the highest conflict scores are selected as the high-conflict group.

We compare the F1 scores produced by ECFMIR and the baseline methods on these high-conflict intent categories within the MintRec2.0 dataset. As presented in Table 3, the best results obtained for each category are highlighted in bold, whereas the second-best results are underlined. Overall, our approach consistently ranks within the top two across all ten high-conflict categories, achieving eight first-place scores and two second-place scores. This superiority stems from the fine-grained confidence weighting scheme employed by our model for unimodal features and the effective integration of cross-modal information, enabling conflicting samples to be more accurately disambiguated. These results demonstrate that ECFMIR significantly improves the recognition robustness of multimodal fusion for high-conflict samples.

cross-modal associations Analysis

To conduct an in-depth analysis of the confidence behavior exhibited by our model under different modality combinations, we plot the confidence discrepancy density maps

Methods	Doubt	Inform	Taunt	Emphasize	Care	Ask for opinions	Ask for help	Praise	Plan	Prevent
MulT	62.00	53.76	16.16	8.7	53.49	55.45	57.14	76.44	47.5	70.18
MAG-BERT	58.54	48.67	22.43	9.3	48.42	53.76	63.16	72.73	48.48	53.57
MISA	58.45	52.04	21.31	7.69	54.32	52.00	63.01	74.42	**63.41**	66.67
SDIF	60.00	55.65	22.61	34.48	**56.31**	59.34	60.00	72.34	58.74	63.01
TCL-MAP	41.48	47.00	18.80	5.71	23.88	52.54	52.5	50.70	53.33	47.76
MVCL-DAF	61.13	54.01	15.56	8.89	53.76	50.00	58.82	73.59	55.42	65.57
ECFMIR	**63.28**	**56.63**	**43.33**	**65.26**	56.00	**64.81**	**68.35**	**80.36**	59.34	**74.19**

Table 3: Performance achieved by different methods on various intent categories.

Methods	ECE (%)	ACE (%)	OE (%)
MAG-BERT	18.24	18.26	24.14
MISA	16.20	**15.56**	30.72
MulT	18.19	18.63	21.39
TCL-MAP	17.61	18.01	22.19
MVCL-DAF	17.35	17.63	26.29
SDIF	16.75	16.68	24.26
ECFMIR	**15.99**	15.82	**20.92**

Table 4: Comparison among the calibration error metrics produced by different methods.

produced for three bimodal pairs (TA, TV, and AV), as illustrated in Figure 4. In each subplot, the horizontal axis represents the confidence difference between the two unimodal features–quantifying modality inconsistency–while the vertical axis denotes the confidence score of the combined modality. The color intensity reflects the sample density distribution. Notably, the TA pair exhibits significantly higher confidence when the modality difference is small, indicating a strong synergistic enhancement effect. The density of the TV pair is concentrated near zero, suggesting a high degree of agreement between these modalities. Conversely, the AV pair generally has lower confidence scores and heightened sensitivity to modality discrepancies, reflecting a more conservative stance in conflicting scenarios. Overall, the model demonstrates stronger confidence under high levels of modality consistency and appropriately attenuates its confidence when conflicts arise, showcasing robust and nuanced uncertainty modeling capabilities.

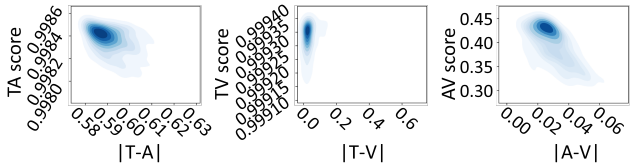

Figure 4: Density plots of the confidence discrepancies produced for different modality pairs (TA, TV, and AV) on the MintRec dataset. The x-axis shows the confidence differences among the unimodal features, reflecting modality inconsistency; the y-axis represents the cross-modal associations confidence score.

Complexity Comparison

We evaluate the complexity of ECFMIR by analyzing its number of parameters, its training and testing times per batch, and its GPU memory consumption. As summarized in Table 5, we compare ECFMIR with other methods on the MintRec dataset. The results demonstrate that ECFMIR achieves competitive performance with a moderate parameter count, a minimal inference time of 9.97 ms per batch, and reduced GPU memory usage.

Method	Parameters	Inference time	Memory
MAG-BERT	112.51 M	10.00 ms	2609.78 MB
MulT	136.15 M	60.00 ms	8664.74 MB
MISA	139.96 M	13.00 ms	3219.90 MB
SDIF	140.63 M	11.46 ms	3336.87 MB
TCL-MAP	270.51 M	17.94 ms	6215.70 MB
MVCL-DAF	241.55 M	16.94 ms	5746.02 MB
ECFMIR	151.65 M	9.97 ms	3414.35 MB

Table 5: Complexity comparison among different methods.

Conclusion

To address the challenge of reliability modeling in multimodal intent recognition, we propose ECFMIR. This method quantifies the reliability of different modalities and their interrelations through a unified confidence lens, effectively mitigating uncertainty and reducing the risk of erroneous predictions for conflicting samples. Evaluations on the MIntRec and MIntRec2.0 benchmarks demonstrate that our method outperforms state-of-the-art techniques. Ablation studies confirm the effectiveness of the model's key components. Further analysis reveals that when significant modality conflicts arise or quality is poor, the model dynamically focuses on more reliable sources. This showcases strong selection capabilities and robustness, particularly in low-resource settings.

References

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

Bhattacharjee, D.; Zhang, T.; Süsstrunk, S.; and Salzmann, M. 2022. Mult: An end-to-end multitask learning trans-

former. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12031–12041.

Chen, Q.; Zhuo, Z.; and Wang, W. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

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

He, C.; Zhang, X.; Song, D.; Shen, Y.; Mao, C.; Wen, H.; Zhu, D.; and Cai, L. 2024. Mixture of attention variants for modal fusion in multi-modal sentiment analysis. *Big Data and Cognitive Computing*, 8(2): 14.

Hu, B.; Zhang, K.; Zhang, Y.; and Ye, Y. 2025. Adaptive Multimodal Fusion: Dynamic Attention Allocation for Intent Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17267–17275.

Huang, S.; Qin, L.; Wang, B.; Tu, G.; and Xu, R. 2024. Sdif-da: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10206–10210. IEEE.

Liu, Y.; Yuan, Z.; Mao, H.; Liang, Z.; Yang, W.; Qiu, Y.; Cheng, T.; Li, X.; Xu, H.; and Gao, K. 2022. 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*, 247–258.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, H.; Zhang, Q.; Zhang, C.; Wu, B.; Fu, H.; Zhou, J. T.; and Hu, Q. 2023. Calibrating multimodal learning. In *International Conference on Machine Learning*, 23429–23450. PMLR.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2020, 2359.

Sun, K.; Xie, Z.; Ye, M.; and Zhang, H. 2024. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26963–26973.

Tellamekala, M. K.; Amiriparian, S.; Schuller, B. W.; André, E.; Giesbrecht, T.; and Valstar, M. 2023. COLD fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 805–822.

Tian, J.; Cheung, W.; Glaser, N.; Liu, Y.-C.; and Kira, Z. 2020. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 5716–5723. IEEE.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Xie, Z.; Yang, Y.; Wang, J.; Liu, X.; and Li, X. 2024. Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 7657–7670.

Zhang, H.; Wang, X.; Xu, H.; Zhou, Q.; Gao, K.; Su, J.; Li, W.; Chen, Y.; et al. 2024a. Mintrec2. 0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. *arXiv preprint arXiv:2403.10943*.

Zhang, H.; Xu, H.; Wang, X.; Zhou, Q.; Zhao, S.; and Teng, J. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM international conference on multimedia*, 1688–1697.

Zhang, L.; Yu, J.; Zhang, S.; Li, L.; Zhong, Y.; Liang, G.; Yan, Y.; Ma, Q.; Weng, F.; Pan, F.; et al. 2024b. Unveiling the impact of multi-modal interactions on user engagement: A comprehensive evaluation in ai-driven conversations. *arXiv preprint arXiv:2406.15000*.

Zhao, F.; Zhang, C.; and Geng, B. 2024. Deep multimodal data fusion. *ACM computing surveys*, 56(9): 1–36.

Zhao, J.; and Li, S. 2024. Evidence modeling for reliability learning and interpretable decision-making under multi-modality medical image segmentation. *Computerized Medical Imaging and Graphics*, 116: 102422.

Zheng, X.; Tang, C.; Wan, Z.; Hu, C.; and Zhang, W. 2023. Multi-level confidence learning for trustworthy multimodal classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11381–11389.

Zhou, Q.; Xu, H.; Li, H.; Zhang, H.; Zhang, X.; Wang, Y.; and Gao, K. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 17114–17122.

Zhu, Z.; Cheng, X.; Chen, Z.; Chen, Y.; Zhang, Y.; Wu, X.; Zheng, Y.; and Xing, B. 2024. InMu-Net: advancing multimodal intent detection via information bottleneck and multisensory processing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 515–524.

Zou, X.; Tang, C.; Zheng, X.; Li, Z.; He, X.; An, S.; and Liu, X. 2023. Dpnet: Dynamic poly-attention network for trustworthy multi-modal classification. In *Proceedings of the 31st ACM international conference on multimedia*, 3550–3559.