RPE-MulT: Reliable Fusion and Prototype Discrimination for Multimodal Intent Recognition

Anonymous ACL submission

Abstract

Multimodal intent recognition aims to infer user intent from textual, acoustic, and visual signals. However, acoustic and visual modalities are often unreliable, and fine-grained intent categories usually have ambiguous boundaries. We propose RPE-MulT, a Reliability-aware Prototype-Enhanced Multimodal Transformer that improves intent prediction from both the fusion and discrimination stages. At the fusion stage, RPE-MulT estimates token-level reliability scores for acoustic and visual tokens and incorporates them into cross-modal attention to suppress unreliable evidence. At the discrimination stage, it adopts a decoupled prototype-enhanced design, where a standard classifier performs final prediction while an auxiliary prototype head uses learnable intent prototypes to impose class-level structural regularization. Comprehensive experiments on two challenging benchmark datasets show that RPE-MulT outperforms strong baselines on accuracy, weighted F1-score, and weighted precision, while ablation, robustness, and representation analyses further verify the effectiveness of the proposed framework.

1 Introduction

Multimodal Intent Recognition (MIR) aims to infer user intent from textual, acoustic, and visual cues, and is fundamental to dialogue systems, virtual assistants, and embodied multimodal agents. Compared with text-only intent recognition, MIR provides richer behavioral evidence, but also introduces two challenges: non-textual modalities are often unreliable, and fine-grained intent categories may have ambiguous decision boundaries.

Existing MIR methods have improved multimodal fusion through cross-modal attention, nonverbal feature injection, modality disentanglement, adaptive fusion, and confidence-aware modeling (Tsai et al., 2019; Rahman et al., 2020; Hazarika et al., 2020; Zhou et al., 2024; Hu et al., 2025; Liu et al., 2026). Despite their effectiveness, two important issues remain underexplored, as illustrated in Fig. 1.

First, acoustic and visual signals are often affected by background noise, silence, occlusion, motion blur, or weak facial expressions. As shown in Fig. 1(a), if such unreliable audio-visual tokens are directly involved in cross-modal attention, they may be over-attended and amplified in the fused intent representation. Although recent reliability-aware methods estimate modality, view, or sample-level confidence (YANG et al., 2025; Hu et al., 2025; Liu et al., 2026), such coarse-grained modeling may be insufficient to suppress individual unreliable source tokens during token-level cross-modal interaction.

Second, existing MIR models are mainly optimized by instance-level classification objectives, which may not explicitly enforce a well-structured intent representation space. As shown in Fig. 1(b), samples from the same intent may be scattered, while different fine-grained intents may overlap. Prototype-based learning offers a natural way to impose class-level structure by encouraging samples to align with their corresponding class prototypes (Li et al., 2021; Pahde et al., 2021; Huang et al., 2025). However, existing prototype-enhanced methods rarely consider how class-level structural regularization can be decoupled from final prediction and coordinated with reliability-aware cross-modal fusion in MIR.

To address these issues, we propose RPE-MulT, a Reliability-aware Prototype-Enhanced Multimodal Transformer for MIR. RPE-MulT improves MIR from two complementary stages. At the fusion stage, built upon MulT (Tsai et al., 2019), it estimates token-level reliability scores for acoustic and visual tokens and injects them into cross-modal attention to suppress unreliable source evidence before multimodal representations are formed. At the discrimination stage, it introduces learnable intent

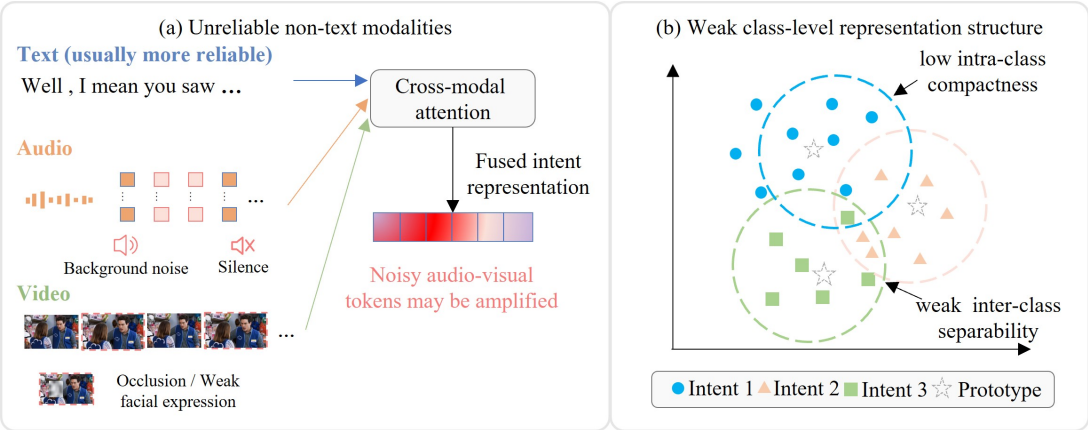

Figure 1: Illustration of the two main challenges in MIR: unreliable non-textual modalities and weak class-level representation structure.

prototypes to regularize fused utterance representations, encouraging intra-class compactness and inter-class separability. In this way, RPE-MulT jointly improves the robustness of multimodal fusion and the discriminability of the representations.

The main contributions are as follows:

1. We propose a reliability-aware fusion strategy that estimates token-level reliability for acoustic and visual cues and incorporates the reliability scores into cross-modal attention to suppress unreliable source evidence.

2. We introduce a decoupled prototype-enhanced discrimination module, where a standard classifier performs final intent prediction while an auxiliary prototype head regularizes the fused representation with learnable intent prototypes, improving intra-class compactness and inter-class separability.

3. We conduct extensive experiments on two challenging benchmark datasets, including comparisons with strong baselines, ablation studies, robustness evaluation under corrupted modalities, and representation analysis.

2 Related Work

2.1 Multimodal Intent Recognition

Multimodal intent recognition extends text-only intent classification by incorporating acoustic and visual cues. Early studies mainly focus on multimodal fusion architectures. MulT (Tsai et al., 2019) models cross-modal interactions through directional cross-modal attention, while MAG-BERT (Rahman et al., 2020) injects nonverbal information into pretrained language representations. MISA (Hazarika et al., 2020) further disentangles modality-invariant and modality-specific factors to reduce modality gaps.

Recent MIR methods explore more adaptive and semantically enriched multimodal modeling. TCL-MAP (Zhou et al., 2024) introduces modality-aware prompting and token-level contrastive learning, while MVCL-DAF (Hu et al., 2025) dynamically allocates intra- and inter-modal attention. CR-3WD (Li et al., 2026) and ECFMIR (Liu et al., 2026) further investigate complementary reasoning and confidence-aware prediction. Different from these methods, RPE-MulT explicitly regulates token-level reliability at the fusion stage and further enhances class-level representation structure at the discrimination stage with learnable intent prototypes.

2.2 Reliability-Aware Multimodal Learning

Reliability-aware modeling is important when multimodal signals are noisy, weak, or incomplete. In MIR, TMIR (YANG et al., 2025) estimates view credibility through subjective logic, while ECFMIR (Liu et al., 2026) studies confidence-focused intent recognition. MVCL-DAF (Hu et al., 2025) and CR-3WD (Li et al., 2026) also address sample-dependent modality quality through adaptive fusion or complementary reasoning. Beyond MIR, PRLF (Bao et al., 2026) shows the effectiveness of confidence estimation in incomplete multimodal learning.

Most existing methods estimate reliability at the modality, view, or sample level. However, cross-modal attention operates at the token level, where individual noisy source tokens may still receive high attention weights. RPE-MulT differs by esti-

mating token-level reliability for acoustic and visual sequences and using the estimated scores to modulate both attention logits and source value aggregation.

2.3 Prototype-Enhanced Representation Learning

Prototype-based learning improves representation structure by aligning samples with class-level semantic centers. PCL (Li et al., 2021) introduces prototypes to capture latent semantic structures in contrastive learning, while Multimodal Prototypical Networks (Pahde et al., 2021) exploit auxiliary textual information for few-shot multimodal classification. In MIR-related studies, CorrKD (Li et al., 2024) uses prototype-based distillation to recover missing semantics, and MVCL-DAF++ (Huang et al., 2025) introduces prototype-aware contrastive alignment.

Unlike prior prototype-enhanced methods that mainly focus on few-shot recognition, contrastive alignment, or distillation, RPE-MulT uses learnable intent prototypes as an auxiliary class-level regularizer for fused utterance representations. This enables prototype learning to complement fusion-stage reliability modeling by improving late-stage intra-class compactness and inter-class separability in MIR.

3 Method

To address unreliable modalities and weak class-level representation structure, we propose RPE-MulT, a Reliability-aware Prototype-Enhanced Multimodal Transformer for MIR. RPE-MulT improves MIR from two complementary stages: reliability-aware fusion and prototype-enhanced discrimination. At the fusion stage, it estimates token-level reliability scores for acoustic and visual cues and incorporates them into cross-modal attention to suppress unreliable source evidence. At the discrimination stage, it adopts a decoupled design, where a standard classifier performs final prediction and an auxiliary prototype head regularizes fused utterance representations with learnable intent prototypes. The overall architecture is shown in Fig. 2.

3.1 Problem Formulation

Given a multimodal utterance, MIR aims to predict its intent label from textual, acoustic, and visual information. We denote the three input modalities as

$$X^t \in \mathbb{R}^{T_t \times d_t}, \quad X^a \in \mathbb{R}^{T_a \times d_a}, \quad X^v \in \mathbb{R}^{T_v \times d_v}, \tag{1}$$

where X^t, X^a, and X^v represent textual, acoustic, and visual sequences, respectively. T_m and d_m denote the sequence length and feature dimension of modality $m \in \{t, a, v\}$. The goal is to learn a prediction function

$$\hat{y} = f(X^t, X^a, X^v), \tag{2}$$

where $\hat{y} \in \{1, \ldots, K\}$ is the predicted intent label and K is the number of intent classes.

3.2 Multimodal Encoding and Fusion

Since the three modalities have different feature dimensions, we first project them into a shared latent space:

$$H^m = X^m W_p^m + b_p^m, \quad m \in \{t, a, v\}, \tag{3}$$

where $W_p^m \in \mathbb{R}^{d_m \times d}$ and $b_p^m \in \mathbb{R}^d$ are learnable parameters, and d is the unified hidden dimension. Positional encodings are added to preserve temporal order.

Following MulT, we perform directional cross-modal attention for each target-source modality pair. For a target modality x and a source modality y, the query, key, and value representations are computed as

$$Q^x = H^x W_Q, \quad K^y = H^y W_K, \quad V^y = H^y W_V, \tag{4}$$

where W_Q, W_K, and W_V are learnable projection matrices.

After obtaining the directional cross-modal representations, we aggregate them for each target modality using modality-specific memory transformers:

$$\begin{aligned} \tilde{H}^t &= \mathrm{Trans}^t_{\mathrm{mem}}\big([Z^{t\leftarrow a}; Z^{t\leftarrow v}]\big), \\ \tilde{H}^a &= \mathrm{Trans}^a_{\mathrm{mem}}\big([Z^{a\leftarrow t}; Z^{a\leftarrow v}]\big), \\ \tilde{H}^v &= \mathrm{Trans}^v_{\mathrm{mem}}\big([Z^{v\leftarrow t}; Z^{v\leftarrow a}]\big), \end{aligned} \tag{5}$$

where $[\cdot;\cdot]$ denotes feature concatenation and $Z^{x\leftarrow y}$ denotes the cross-modal representation from source modality y to target modality x.

We obtain utterance-level summaries by pooling the memory outputs:

$$u^t = \mathrm{Pool}(\tilde{H}^t), \quad u^a = \mathrm{Pool}(\tilde{H}^a), \quad u^v = \mathrm{Pool}(\tilde{H}^v). \tag{6}$$

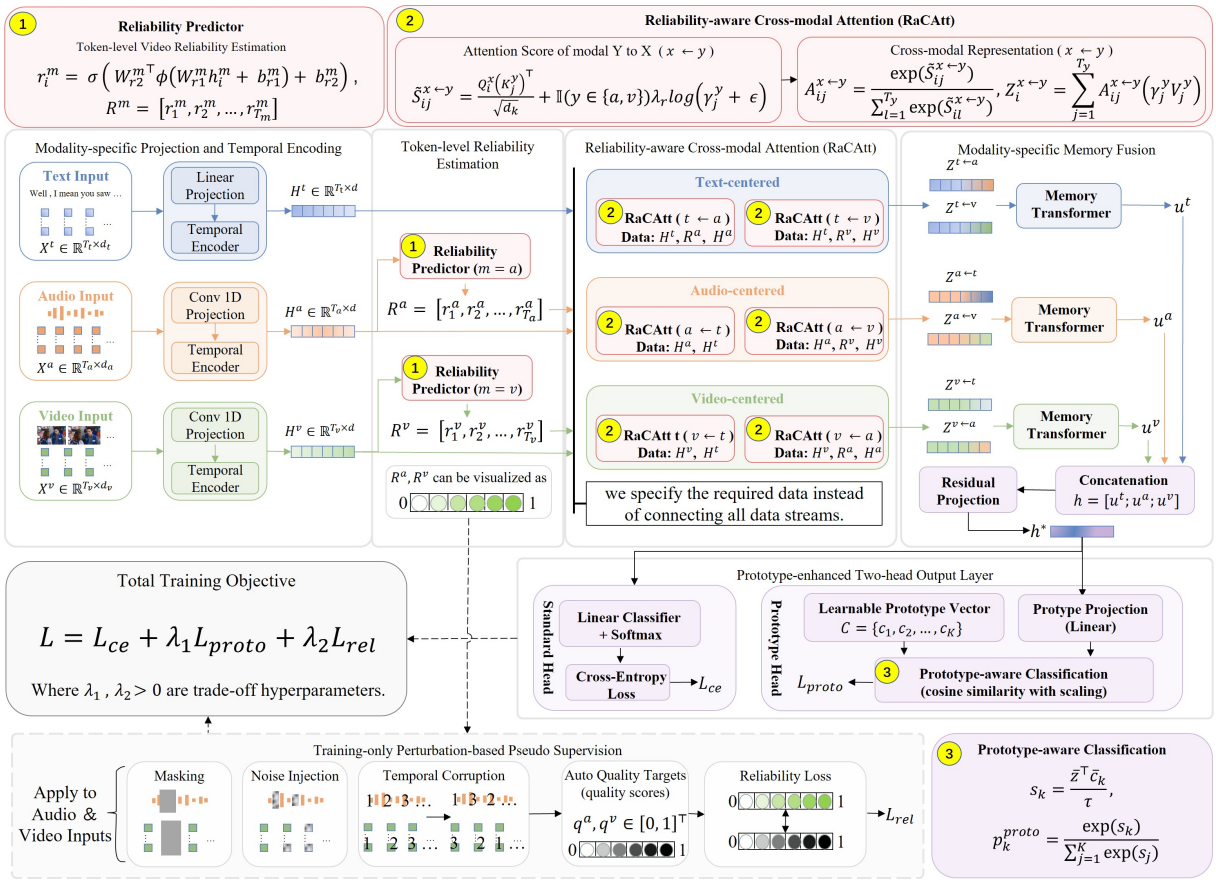

Figure 2: Overview of the proposed RPE-MulT. Textual, acoustic, and visual features are first projected into a shared latent space. Reliability-aware cross-modal attention suppresses unreliable audio-visual source tokens during multimodal interaction. The fused utterance representation is then optimized by both a standard classifier and a prototype-based auxiliary head.

The final multimodal representation is obtained by concatenating the three modality summaries:

$$h = [u^t; u^a; u^v]. \tag{7}$$

A residual projection layer is further applied:

$$h^* = W_2 \mathrm{Dropout}\big(\mathrm{ReLU}(W_1 h)\big) + h, \tag{8}$$

where h^* is the final fused utterance representation.

3.3 Reliability-aware Cross-modal Attention

In standard cross-modal attention, source tokens are weighted mainly according to content similarity. However, acoustic and visual tokens may be unreliable due to noise, silence, occlusion, or weak expressions. To reduce negative transfer from such unreliable evidence, we introduce token-level reliability estimation for acoustic and visual modalities.

For the i-th token of modality $m \in \{a, v\}$, its reliability score is computed as

$$r_i^m = \sigma\left(w_{r2}^{m\top} \phi\left(W_{r1}^m h_i^m + b_{r1}^m\right) + b_{r2}^m\right), \tag{9}$$

where $\phi(\cdot)$ is a nonlinear activation function and $\sigma(\cdot)$ is the sigmoid function. Thus, $r_i^m \in (0, 1)$, where a larger value indicates a more reliable token.

Since explicit token-level quality annotations are unavailable, we train the reliability estimator with perturbation-based pseudo supervision. During training, we randomly corrupt acoustic and visual sequences with perturbations such as temporal masking and Gaussian noise. For each token, we define a pseudo corruption intensity $\eta_i^m \in [0, 1]$, where a larger value indicates stronger corruption. The corresponding pseudo reliability target is

$$q_i^m = 1 - \eta_i^m. \tag{10}$$

For uncorrupted tokens, we set $q_i^m = 1$. When a perturbation is applied uniformly to an entire sequence, η_i^m is set to the sequence-level corruption intensity ρ^m.

The reliability loss is formulated as

$$L_{\mathrm{rel}} = \sum_{m \in \{a,v\}} \frac{1}{T_m} \sum_{i=1}^{T_m} \|r_i^m - q_i^m\|_2^2. \tag{11}$$

For cross-modal attention, the standard attention score between the i-th target token and the j-th source token is

$$S_{ij}^{x\leftarrow y} = \frac{Q_i^x (K_j^y)^\top}{\sqrt{d_k}}. \tag{12}$$

We define a source-dependent reliability coefficient:

$$\gamma_j^y = \begin{cases} r_j^y, & y \in \{a, v\}, \\ 1, & y = t. \end{cases} \tag{13}$$

The reliability coefficient is injected into the attention logits as a reliability prior:

$$\begin{aligned} \tilde{S}_{ij}^{x\leftarrow y} =& S_{ij}^{x\leftarrow y} \\ &+ \mathbb{I}(y \in \{a, v\}) \lambda_r \log(\gamma_j^y + \epsilon), \end{aligned} \tag{14}$$

where $\mathbb{I}(\cdot)$ is the indicator function, λ_r controls the strength of reliability modulation, and ϵ is a small constant for numerical stability.

The normalized attention weight is computed as

$$A_{ij}^{x\leftarrow y} = \frac{\exp(\tilde{S}_{ij}^{x\leftarrow y})}{\sum_{l=1}^{T_y} \exp(\tilde{S}_{il}^{x\leftarrow y})}. \tag{15}$$

The reliability-aware cross-modal output is then obtained by

$$Z_i^{x\leftarrow y} = \sum_{j=1}^{T_y} A_{ij}^{x\leftarrow y} \left(\gamma_j^y V_j^y \right). \tag{16}$$

Adding $\log(\gamma_j^y + \epsilon)$ to the attention logits uses reliability as a prior to adjust the relative attention distribution over source tokens, while multiplying γ_j^y with the value vectors directly controls the amount of information propagated from each source token. Thus, unreliable acoustic and visual tokens are suppressed both during attention selection and feature aggregation.

3.4 Prototype-enhanced Intent Discrimination

Although reliability-aware fusion improves the quality of multimodal interaction, the fused representation may still lack explicit class-level structure. To improve representation discriminability without over-constraining the main fused representation, we introduce a prototype-enhanced intent discrimination module with a decoupled prototype space.

Assume there are K intent classes. We define one learnable prototype for each class:

$$C = \{c_1, c_2, \ldots, c_K\}, \quad c_k \in \mathbb{R}^{d_z}, \tag{17}$$

where d_z is the prototype-space dimension. The fused representation h^* is first projected into the prototype space:

$$z = W_z h^* + b_z. \tag{18}$$

Both the projected representation and class prototypes are L_2-normalized:

$$\bar{z} = \frac{z}{\|z\|_2 + \epsilon}, \quad \bar{c}_k = \frac{c_k}{\|c_k\|_2 + \epsilon}. \tag{19}$$

The similarity between $\bar{z}$ and the k-th class prototype is computed by temperature-scaled cosine similarity:

$$s_k = \frac{\bar{z}^\top \bar{c}_k}{\tau}, \tag{20}$$

where τ is the temperature parameter.

The prototype-based probability is defined as

$$p_k^{\text{proto}} = \frac{\exp(s_k)}{\sum_{j=1}^{K} \exp(s_j)}. \tag{21}$$

The prototype loss is

$$L_{\text{proto}} = -\log p_y^{\text{proto}}, \tag{22}$$

where y is the ground-truth label. This objective serves as an auxiliary class-level regularizer that promotes intra-class compactness and inter-class separability in the projected prototype space.

To stabilize optimization, we additionally keep a standard linear classifier as the primary prediction head. Different from the prototype head, this classifier directly operates on the fused utterance representation h^*:

$$p^{\text{cls}} = \text{softmax}(W_c h^* + b_c). \tag{23}$$

The standard classification loss is

$$L_{\text{ce}} = -\log p_y^{\text{cls}}. \tag{24}$$

During inference, we use the standard classifier output p^{cls} for intent prediction, while the prototype head is used as an auxiliary training objective.

We deliberately decouple the primary classification space and the auxiliary prototype space. The standard classifier operates on h^*, while the prototype head projects h^* into a dedicated metric space z for cosine-based prototype discrimination. This design allows h^* to preserve rich multimodal information for final prediction, while z focuses on class-level structural regularization. Since z is obtained from h^* through a learnable projection, the prototype loss still back-propagates structural supervision to h^* without over-constraining the main representation space.

3.5 Training Objective

The final training objective is defined as

$$L = L_{\text{ce}} + \lambda_1 L_{\text{proto}} + \lambda_2 L_{\text{rel}}, \tag{25}$$

where λ_1 and λ_2 are hyperparameters that balance prototype-enhanced supervision and reliability supervision.

The classification loss optimizes the main intent prediction task, the prototype loss improves class-level representation structure, and the reliability loss encourages quality-aware estimation of acoustic and visual tokens. Therefore, RPE-MulT jointly improves the reliability of early multimodal fusion and the discriminability of late intent prediction.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed method on two challenging benchmarks: MIntRec (Zhang et al., 2022) and MIntRec2.0 (Zhang et al., 2024). MIntRec contains 2,224 multimodal samples, covering 20 intent categories. MIntRec2.0 is a more challenging benchmark for MIR. It contains 15,040 multimodal samples, and defines 30 fine-grained in-scope intent classes. See details in Appendix A.1.

Evaluation Metrics. Following previous MIR studies, we report Accuracy (ACC), Weighted F1-score (WF1), Weighted Precision (WP), and macro-averaged Recall (R). All results are averaged over four runs with random seeds $\{1, 2, 3, 4\}$.

Baselines. We compare RPE-MulT with representative baselines, including MulT (Tsai et al., 2019), MAG-BERT (Rahman et al., 2020), MISA (Hazarika et al., 2020), TCL-MAP (Zhou et al., 2024), MVCL-DAF (Hu et al., 2025), and ECFMIR (Liu et al., 2026). Baseline descriptions are provided in Appendix A.2.

Details. All methods are evaluated under the same data split, feature setting, and evaluation protocol. See details in Appendix A.

4.2 Main Results

Table 1 reports the main results on MIntRec and MIntRec2.0. On MIntRec, RPE-MulT achieves the best performance across all metrics, obtaining 76.35% ACC, 75.72% WF1, 75.99% WP, and 74.83% Recall. Compared with the strongest baseline ECFMIR, RPE-MulT improves ACC, WF1, WP, and Recall by 1.77, 1.31, 0.94, and 0.71 absolute points, respectively. These results show that the proposed fusion-stage reliability modeling and discrimination-stage prototype regularization are effective on the MIntRec benchmark.

On MIntRec2.0, RPE-MulT also achieves the best ACC, WF1, and WP, reaching 61.03%, 59.76%, and 59.98%, respectively. Compared with ECFMIR, RPE-MulT improves ACC, WF1, and WP by 1.51, 1.14, and 1.31 points. Although ECFMIR obtains the highest Recall, RPE-MulT achieves comparable Recall while delivering stronger overall performance on the other three metrics. This indicates that RPE-MulT produces more accurate and precise predictions for fine-grained multimodal intent recognition.

To further evaluate the stability of different methods, we report the mean and standard deviation over four random seeds in Appendix B. The results show that RPE-MulT consistently achieves strong performance across different runs. It is also observed that all methods achieve substantially lower performance on MIntRec2.0 than on MIntRec. This confirms that MIntRec2.0 is more challenging due to its larger scale and more fine-grained intent taxonomy. Therefore, to provide a stricter evaluation of the proposed modules, the following ablation study, robustness analysis, and prototype representation visualization are conducted on MIntRec2.0.

4.3 Ablation Study

To verify the contribution of each component, we conduct ablation experiments. The variant w/o Relia. removes the reliability-aware module, w/o Proto. removes the prototype-enhanced module, and w/o Relia. & Proto. removes both modules.

Table 2 shows the contribution of each component. Removing the reliability-aware module decreases ACC, WF1, WP, and Recall by 1.22, 1.21, 0.61, and 1.42 points, respectively, confirming the importance of fusion-stage reliability modeling. Removing the prototype-enhanced module causes smaller but consistent drops on ACC, WF1, and Recall, indicating that discrimination-stage prototype supervision improves the discriminability of fused intent representations.

When both modules are removed, the model obtains the weakest performance, with drops of 2.69 ACC, 1.95 WF1, 4.76 WP, and 4.02 Recall compared with the full RPE-MulT. These results demonstrate that the two stages are complementary: reliability-aware fusion improves the quality of

Table 1: Main results on MIntRec and MIntRec2.0. The best results are highlighted in bold.

Method	MIntRec				MIntRec2.0			
	ACC(%)	WF1(%)	WP(%)	R(%)	ACC(%)	WF1(%)	WP(%)	R(%)
MulT	72.52	71.80	72.60	67.44	57.55	55.52	55.91	48.13
MAG-BERT	68.99	68.81	69.26	63.80	57.70	55.40	56.25	47.75
MISA	72.36	72.64	73.72	70.61	58.58	56.64	57.03	48.71
TCL-MAP	73.69	73.38	73.90	72.59	58.88	57.32	57.60	50.75
MVCL-DAF	73.01	72.02	72.14	71.01	59.27	57.53	58.26	50.19
ECFMIR	74.58	74.41	75.05	74.12	59.52	58.62	58.67	**54.96**
RPE-MulT (ours)	**76.35**	**75.72**	**75.99**	**74.83**	**61.03**	**59.76**	**59.98**	54.17

Table 2: Ablation study of RPE-MulT on MIntRec2.0.

Method / Variant	ACC(%)	WF1(%)	WP(%)	R(%)
RPE-MulT	**61.03**	**59.76**	59.98	**54.17**
w/o Relia.	59.81	58.55	59.37	52.75
w/o Proto.	60.45	59.42	**60.36**	53.98
w/o Relia. & Proto.	58.34	57.81	55.22	50.15

multimodal interaction, while prototype-enhanced discrimination regularizes the resulting intent representation. Although w/o Proto. slightly improves WP, the full model achieves the best ACC, WF1, and Recall, showing the best overall balance.

4.4 Robustness to Unreliable Modalities

To evaluate robustness under realistic input degradation, we corrupt all input modalities during evaluation, including textual features to simulate automatic speech recognition (ASR) or transcription errors, with corruption probabilities $\rho^m \in \{10\%, 20\%, 50\%\}$. We report results using the validation-best checkpoint and compare RPE-MulT with w/o Relia., w/o Proto., and MulT. The WF1 degradation is defined as:

$$\Delta\text{WF1} = \text{WF1}_{\text{clean}} - \text{WF1}_{\text{corrupted}}. \quad (26)$$

As shown in Table 3, RPE-MulT consistently achieves the best ACC and WF1 under both clean and corrupted settings. Under the severe 50% corruption setting, RPE-MulT obtains 54.21% ACC and 54.18% WF1, outperforming w/o Relia. by 3.20 and 2.64 points and MulT by 5.51 and 5.33 points, respectively. Moreover, removing the reliability-aware module leads to larger WF1 drops across all corruption levels, i.e., 2.57 vs. 1.80 at 10%, 3.78 vs. 2.86 at 20%, and 8.40 vs. 6.21 at 50%. These results verify that fusion-stage reliability modeling improves robustness under degraded multimodal inputs. Although w/o Proto. shows slightly smaller ΔWF1 in some settings, its absolute ACC and WF1 are consistently lower than those of RPE-MulT, indicating that prototype-enhanced discrimination further improves prediction performance.

4.5 Prototype Representation Visualization

To examine whether the prototype-enhanced module improves representation structure, we conduct qualitative and quantitative analyses on selected difficult intent classes. Instead of random selection, we analyze the confusion matrices of RPE-MulT and w/o Proto. and select four representative classes, including *Oppose*, *Explain*, *Doubt*, and *Advise*. These classes are frequently involved in misclassification or exhibit fine-grained semantic ambiguity. The full confusion matrices used for class selection are provided in Appendix C.1.

For a fair comparison, we extract the fused utterance representations h^* before the final classifier from both RPE-MulT and w/o Proto. The representations from the two models are concatenated and projected into a shared two-dimensional space using UMAP(McInnes et al., 2018). As shown in Fig. 3, w/o Proto. produces more scattered and overlapped distributions, while RPE-MulT forms more compact and separable class structures.

Since UMAP visualization can be affected by the projection algorithm, we further compute quantitative separability metrics using the original high-dimensional representations before projection, including intra-class distance, inter-class distance (Fisher, 1936), inter/intra ratio, silhouette score (Rousseeuw, 1987), and Davies–Bouldin Index (DBI) (Davies and Bouldin, 1979). The de-

Table 3: Robustness results under unreliable modalities with different corruption probabilities (ρ^m) on MIntRec2.0. ΔWF1(%) denotes the WF1 decrease compared with the clean setting. Bold indicates the best ACC(%) or WF1(%), and underline indicates the smallest WF1 drop.

Method	Clean		10%			20%			50%		
	ACC	WF1	ACC	WF1	ΔWF1	ACC	WF1	ΔWF1	ACC	WF1	ΔWF1
RPE-MulT	**61.39**	**60.39**	**59.47**	**58.59**	1.80	**58.19**	**57.53**	2.86	**54.21**	**54.18**	6.21
w/o Relia.	61.09	59.94	58.63	57.37	2.57	57.06	56.16	3.78	51.01	51.54	8.40
w/o Proto.	60.55	59.52	**59.47**	58.45	<u>1.07</u>	57.99	57.16	<u>2.36</u>	53.42	53.68	<u>5.84</u>
MulT	58.96	57.81	56.91	55.89	1.92	55.29	54.48	3.33	48.70	48.85	8.96

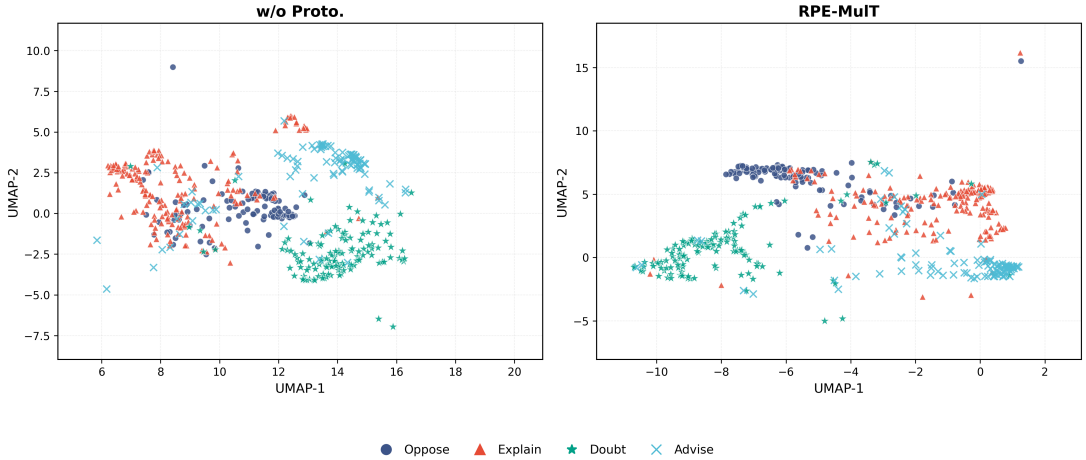

Figure 3: UMAP visualization of fused utterance representations h^* for selected difficult intent classes. The classes are selected according to the confusion matrix analysis. Compared with w/o Proto., RPE-MulT produces more compact and separable class distributions.

Table 4: Quantitative analysis of representation compactness and separability on MIntRec2.0. Metrics are computed using high-dimensional fused representations before UMAP projection.

Method	Intra ↓	Inter ↑	Inter/Intra ↑	Sil. ↑	DBI ↓
w/o Proto.	0.2518	0.4521	1.7953	0.2783	2.0421
RPE-MulT	**0.2337**	**0.5009**	**2.1433**	**0.3278**	**1.7669**

tailed definitions are provided in Appendix C.3.

As shown in Table 4, RPE-MulT consistently improves all separability metrics over w/o Proto. It reduces intra-class distance from 0.2518 to 0.2337 and increases inter-class distance from 0.4521 to 0.5009. The inter/intra ratio and silhouette score also improve, while DBI decreases from 2.0421 to 1.7669. These results confirm that discrimination-stage prototype supervision encourages more compact and discriminative intent representations.

5 Conclusion

In this paper, we propose RPE-MulT, a reliability-aware prototype-enhanced multimodal Transformer for MIR. RPE-MulT improves multimodal intent prediction from two complementary stages. At the fusion stage, it estimates token-level reliability scores for acoustic and visual tokens and incorporates them into cross-modal attention to suppress unreliable source evidence. At the discrimination stage, it introduces a decoupled prototype-enhanced module, where a standard classifier performs final intent prediction while an auxiliary prototype head regularizes fused representations with class-level structural constraints. Experiments on two benchmark datasets show that RPE-MulT outperforms strong MIR baselines on ACC, WF1, and WP. Ablation studies, robustness evaluation under corrupted multimodal inputs, and representation analysis further verify the effectiveness of the proposed modules. These results demonstrate the combined benefits of improving token-level reliability during multimodal fusion and enhancing class-level representation structure during intent discrimination.

Limitations

RPE-MulT still has several limitations. First, although we evaluate the proposed method on two benchmark datasets, MIntRec and MIntRec2.0, both datasets are mainly constructed from scripted or collected conversational video sources. Therefore, the generalization ability of RPE-MulT to more diverse real-world scenarios, languages, domains, and spontaneous human–machine interactions remains to be further investigated.

Second, the reliability-aware module is trained with perturbation-based pseudo supervision rather than human-annotated modality quality labels. Although the robustness experiments show that this strategy improves resistance to corrupted inputs, synthetic perturbations such as temporal masking and Gaussian noise may not fully capture complex real-world degradation, including overlapping speech, severe occlusion, low-resolution videos, inaccurate face detection, or cross-modal asynchrony.

Third, RPE-MulT mainly models the reliability of acoustic and visual modalities while treating textual inputs as relatively reliable. This assumption is reasonable when clean transcriptions are available, but may be suboptimal in practical systems where text is produced by automatic speech recognition or contains transcription errors, informal expressions, or incomplete utterances. Although our robustness evaluation includes corrupted textual inputs to simulate ASR or transcription errors, the current reliability estimator mainly models acoustic and visual reliability. Extending token-level reliability estimation to textual inputs remains future work.

Finally, the current prototype-enhanced module uses one learnable prototype for each intent class. However, fine-grained intents may contain multiple semantic subgroups and diverse multimodal expressions. A single prototype may be insufficient to represent such intra-class diversity, especially for long-tailed or highly ambiguous intent categories. Future work could explore adaptive multi-prototype learning or prototype construction conditioned on dialogue context.

Potential Risks

Although this work is conducted on public benchmark datasets and is intended for research purposes, multimodal intent recognition may introduce potential risks in real-world deployment. Since MIR systems infer user intent from text, speech, and visual behavior, they could be misused for privacy-invasive monitoring, workplace surveillance, or behavioral profiling without informed consent. Practical deployment should therefore require explicit user consent, data minimization, secure storage, and restrictions on secondary use of multimodal recordings.

MIR models may also inherit or amplify social biases. Acoustic and visual signals can vary across languages, accents, cultures, age groups, and recording conditions, leading to unequal performance for underrepresented populations. These concerns are related to broader risks of bias and representational harms in language and multimodal technologies (Sap et al., 2020; Sheng et al., 2021). Subgroup-level evaluation and fairness auditing are necessary before deployment.

In addition, incorrect intent predictions may lead to inappropriate responses in user-facing systems, especially for sensitive intents such as refusal, complaint, warning, or requests for help. RPE-MulT improves robustness to unreliable modalities, but it should still be used as an assistive component rather than an autonomous decision-maker in high-stakes settings. Finally, although our model does not train large foundation models from scratch, repeated experiments and pretrained feature extraction incur computational cost. Future work should explore more efficient and privacy-preserving MIR systems.

References

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

Jindi Bao, Jianjun Qian, Mengkai Yan, and Jian Yang. 2026. Progressive representation learning for multimodal sentiment analysis with incomplete modalities. *arXiv preprint arXiv:2603.09111*.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, and 1 others. 2019. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7:179–188.

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

Bo Hu, Kai Zhang, Yanghai Zhang, and Yuyang Ye. 2025. Adaptive multimodal fusion: dynamic attention allocation for intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17267–17275.

Haofeng Huang, Yifei Han, Long Zhang, Bin Li, and Yangfan He. 2025. Mvcl-daf++: Enhancing multimodal intent recognition via prototype-aware contrastive alignment and coarse-to-fine dynamic attention fusion. *arXiv preprint arXiv:2509.17446*.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. 2021. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468.

Senhao Li, Xiaoliang Chen, Duoqian Miao, Hongyun Zhang, Xiaolin Qin, and Peng Lu. 2026. Bridging modality gaps: Cross-modal complementary learning with three-way decision for multimodal intent recognition. *Expert Systems with Applications*, page 131927.

Yi Liu, Qimeng Yang, and Lanlan Lu. 2026. Who should i trust? explicit confidence-focused multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23891–23898.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. 2021. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2644–2653.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5477–5490.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint*

conference on natural language processing (volume 1: long papers), pages 4275–4293.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6558–6569.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Ying YANG, Yanqiu YANG, and Bengong YU. 2025. Multimodal intent recognition method with view reliability. *Journal of Electronics & Information Technology*, 47(6):1966–1975.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Wenrui Li, Yanting Chen, and 1 others. 2024. Mintrec2. 0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pages 1688–1697.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17114–17122.

Table 5: Overview of the two multimodal intent recognition datasets.

Dataset	Samples	Modalities	Intent Classes
MIntRec	2,224	T/A/V	20
MIntRec2.0	15,040	T/A/V	30

Table 6: Train/validation/test splits used in our experiments.

Dataset	Train	Validation	Test	Total
MIntRec	1,334	445	445	2,224
MIntRec2.0	9,989	1,821	3,230	15,040

A Experiment Details

A.1 Dataset Details

We evaluate RPE-MulT on two benchmark datasets for multimodal intent recognition: MIntRec and MIntRec2.0. Both datasets provide textual, acoustic, and visual information for conversational utterances, enabling the evaluation of multimodal intent recognition models under realistic dialogue scenarios. Table 5 summarizes the overall statistics of the two datasets, and Table 6 reports the data splits used in our experiments.

MIntRec. MIntRec is a benchmark dataset for multimodal intent recognition. It is collected from dialogues in the television series *Superstore* and contains 2,224 utterance-level samples. Each sample is aligned with textual, acoustic, and visual modalities, making it suitable for studying multimodal intent understanding in conversational scenarios. The dataset defines 20 intent classes, which are organized into two high-level intent types: intents expressing emotions or attitudes, and intents aimed at achieving goals. Representative intent labels include *Complain*, *Praise*, *Inform*, and *Advise*. In our experiments, we follow the standard train/validation/test split, approximately 3:1:1, where 1,334 samples are used for training, 445 for validation, and 445 for testing.

MIntRec2.0. MIntRec2.0 extends the original MIntRec dataset by increasing both data scale and intent coverage. It contains 1,245 multi-turn, multi-party dialogues and 15,040 utterance-level samples, where each utterance is associated with text, audio, video, an intent label, and speaker information. Compared with MIntRec, MIntRec2.0 provides richer conversational contexts and a more fine-grained intent taxonomy. It contains 30 in-scope intent classes and naturally occurring out-of-scope samples, including 9,304 in-scope and 5,736 out-of-scope utterances. This design enables more rigorous evaluation under realistic and open-world conversational settings. In our experiments, we use the official split of MIntRec2.0, approximately 7:1:2, with 9,989 training samples, 1,821 validation samples, and 3,230 test samples.

Compared with MIntRec, MIntRec2.0 is substantially more challenging, with nearly seven times more samples, broader intent coverage, finer-grained categories, and naturally occurring out-of-scope utterances. Therefore, while we report main results on both datasets, we conduct further ablation studies, robustness evaluation, and prototype representation analysis on MIntRec2.0 to provide a stricter evaluation of the proposed modules.

A.2 Baseline Descriptions

We compare RPE-MulT with representative multimodal intent recognition baselines. The abbreviations in parentheses after the method name denote the publication venue and year.

MulT. MulT (ACL-19) (Tsai et al., 2019) employs directional cross-modal attention to model dependencies among asynchronous multimodal inputs. It is suitable for multimodal intent recognition because acoustic cues or visual expressions may occur before or after the key textual phrase. As a widely used Transformer-based multimodal fusion architecture, MulT serves as the backbone reference for evaluating the effectiveness of the proposed reliability-aware and prototype-enhanced modules.

MAG-BERT. MAG-BERT (ACL-20) (Rahman et al., 2020) integrates nonverbal information into a BERT-based language encoder through a multimodal adaptation gate. By injecting acoustic and visual features into pretrained textual representations, MAG-BERT provides a strong reference for text-centered multimodal fusion. In multimodal intent recognition, it helps evaluate how much improvement can be obtained by enriching language representations with prosodic and facial cues.

MISA. MISA (MM-20) (Hazarika et al., 2020) learns modality-invariant and modality-specific representations by projecting each modality into separate latent subspaces. This design helps reduce modality gaps while preserving complementary modality information. The learned representations are then fused for downstream prediction. MISA is

included as a representative disentanglement-based multimodal learning method.

TCL-MAP. TCL-MAP (AAAI-24) (Zhou et al., 2024) introduces modality-aware prompting and token-level contrastive learning for multimodal intent recognition. By aligning label semantics with multimodal token representations, TCL-MAP aims to improve fine-grained intent discrimination, especially when discriminative words, acoustic cues, or visual frames are sparse. It serves as a recent prompt-based and contrastive-learning baseline.

MVCL-DAF. MVCL-DAF (AAAI-25) (Hu et al., 2025) combines multi-view contrastive learning with dynamic attention allocation fusion. It treats different modalities as complementary views and uses text and intent labels as anchors to align visual and acoustic representations. Its dynamic attention mechanism adaptively assigns weights to informative modality-specific cues, making it a strong adaptive fusion baseline for multimodal intent recognition.

ECFMIR. ECFMIR (AAAI-26) (Liu et al., 2026) focuses on confidence-aware multimodal intent recognition. It constructs cross-modal associative features and explicitly models their reliability through a confidence estimation mechanism. Specifically, ECFMIR estimates feature uncertainty and transforms it into an unreliability index, which is further converted into a confidence score. This allows unreliable features to contribute less to the final prediction, making ECFMIR a strong reliability-aware baseline.

A.3 Implementation Details

For both MIntRec and MIntRec2.0, we follow the standard multimodal intent recognition setting and extract textual, acoustic, and visual features for each utterance. Since the sequence lengths of different modalities vary across samples, we apply zero-padding to shorter sequences for length normalization.

In the robustness experiment, for textual features, corruption is implemented by temporal masking to simulate missing or erroneous ASR-derived textual tokens. For acoustic and visual features, we apply temporal masking and Gaussian noise to simulate unreliable non-textual signals.

MIntRec. For the MIntRec dataset, we use a pretrained RoBERTa model (Liu et al., 2019) with the Hugging Face Transformers library (Wolf et al., 2020) to extract 1024-dimensional textual features. For the acoustic modality, we employ a pretrained wav2vec 2.0 (Baevski et al., 2020) model implemented in the Hugging Face Transformers library to extract 768-dimensional acoustic features. For the visual modality, we use a pretrained Faster R-CNN with a ResNet-50 backbone from MMDetection (Chen et al., 2019) to extract 256-dimensional visual features. The maximum sequence lengths are set to 30, 480, and 230 for text, audio, and visual features, respectively.

MIntRec2.0. For the MIntRec2.0 dataset, we use RoBERTa with the Hugging Face Transformers library to extract 1024-dimensional textual features. For visual and acoustic modalities, we use a Mask R-CNN from MMDetection to extract 256-dimensional visual features, and a WavLM model (Chen et al., 2022) in Hugging Face Transformers to extract 768-dimensional acoustic features. The maximum sequence lengths are set to 70, 150, and 200 for text, visual, and audio features, respectively.

For RPE-MulT, we set the reliability-aware attention coefficient to $\lambda_r = 1.0$ and the reliability loss weight to $\lambda_2 = 0.1$. To construct pseudo-reliability supervision, each audio or visual sample is corrupted with probability 0.5, where the maximum temporal masking ratio is set to 0.15, and the Gaussian noise standard deviation is set to 0.01. For prototype learning, we set $\lambda_1 = 0.5$, the prototype-space dimension to $d_z = 128$, and the temperature parameter to $\tau = 0.07$.

A.4 Training and Hyperparameters

We train each model for up to 70 epochs and apply early stopping on the validation loss with a patience of 8 epochs. The batch size is set to 16 for all methods. Optimization is performed using AdamW(Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-6}. All experiments are implemented in PyTorch and run on NVIDIA RTX 3090 Ti GPUs.

Table 7 reports the main hyperparameters used for training RPE-MulT. Unless otherwise specified, the same hyperparameter setting is used for both MIntRec and MIntRec2.0.

A.5 Model Size and Budget

To evaluate the additional computational cost introduced by RPE-MulT, we compare its model size and training time with the MulT backbone. As

Table 7: Training hyperparameters.

Hyperparameter	Value
Padding mode	zero padding
Padding location	end
Modality alignment	not required
Training batch size	16
Validation batch size	16
Test batch size	16
Maximum training epochs	70
Early stopping patience	8
Learning rate	5×10^{-6}
Gradient clipping	0.5
Attention mask	True
Destination feature dimension	80
Number of attention heads	4
Number of Transformer layers	8
Text Conv1D kernel size	5
Visual Conv1D kernel size	1
Acoustic Conv1D kernel size	1
Attention dropout	0.0
Visual attention dropout	0.1
Acoustic attention dropout	0.1
ReLU dropout	0.3
Embedding dropout	0.0
Residual dropout	0.0
Output dropout	0.0

Table 8: Model size and training budget. Training time denotes the approximate time for one complete training run under our experimental environment.

Model	Parameters	Extra Parameters
MulT	367.49M	–
RPE-MulT	367.56M	71,970

Table 9: Approximate training time of RPE-MulT on different datasets.

Dataset	Training Time
MIntRec	0.87 h
MIntRec2.0	3.50 h

shown in Table 8, MulT contains approximately 367.49M trainable parameters, while RPE-MulT contains approximately 367.56M trainable parameters. Compared with MulT, RPE-MulT only introduces 71,970 additional trainable parameters, corresponding to about 0.02% relative increase. This indicates that the proposed reliability-aware and prototype-enhanced modules improve performance with negligible parameter overhead.

We also report the approximate training time of RPE-MulT on the two datasets in Table 9. Under our experimental environment on NVIDIA RTX 3090 Ti GPU, training RPE-MulT takes about 0.87 hours on MIntRec and about 3.50 hours on MIntRec2.0. The increased training time on MIntRec2.0 mainly comes from its substantially larger data scale.

B Standard Deviation over Multiple Runs

To evaluate the stability of different methods, we report the mean and standard deviation over four random seeds on MIntRec and MIntRec2.0. The standard deviation is computed across four independent runs with random seeds $\{1, 2, 3, 4\}$. As shown in Table 10, RPE-MulT maintains strong and stable performance on both datasets, consistently achieving the best ACC, WF1, and WP. These results further support the robustness of the proposed reliability-aware fusion and prototype-enhanced discrimination modules.

C Details of Prototype Representation Visualization

C.1 Confusion Matrix and Class Selection

To select representative classes for prototype representation visualization, we analyze the full confusion matrices of RPE-MulT and w/o Proto., as shown in Fig. 4. Instead of simply selecting the classes with the largest number of errors, we choose four representative difficult intent classes, including *Oppose*, *Explain*, *Doubt*, and *Advise*. These classes cover different types of fine-grained semantic ambiguity and correspond to visible off-diagonal confusion patterns in the confusion matrices.

Specifically, *Explain* and *Oppose* form a stable bidirectional confusion pair in both models. In RPE-MulT, *Explain* $\rightarrow$ *Oppose* occurs 13 times and *Oppose* $\rightarrow$ *Explain* occurs 10 times; in w/o Proto., the corresponding confusions are 12 and 10 times, respectively. This indicates that these two intent

Table 10: Mean and standard deviation over four random seeds on MIntRec and MIntRec2.0.

Method	MIntRec				MIntRec2.0			
	ACC(%)	WF1(%)	WP(%)	R(%)	ACC(%)	WF1(%)	WP(%)	R(%)
MulT	72.52 ± 1.10	71.80 ± 0.51	72.60 ± 1.19	67.44 ± 0.45	57.55 ± 0.87	55.52 ± 0.72	55.91 ± 0.73	48.13 ± 0.56
MAG-BERT	68.99 ± 1.03	68.81 ± 0.90	69.26 ± 0.45	63.80 ± 0.86	57.70 ± 1.04	55.40 ± 1.13	56.25 ± 1.12	47.75 ± 0.94
MISA	72.36 ± 0.81	72.64 ± 0.54	73.72 ± 1.05	70.61 ± 0.57	58.58 ± 0.77	56.64 ± 0.72	57.03 ± 1.11	48.71 ± 1.02
TCL-MAP	73.69 ± 0.94	73.38 ± 1.06	73.90 ± 0.52	72.59 ± 1.04	58.88 ± 0.56	57.32 ± 0.52	57.60 ± 1.20	50.75 ± 1.09
MVCL-DAF	73.01 ± 1.15	72.02 ± 1.17	72.14 ± 0.99	71.01 ± 0.68	59.27 ± 1.09	57.53 ± 1.10	58.26 ± 1.11	50.19 ± 0.64
ECFMIR	74.58 ± 0.62	74.41 ± 1.29	75.05 ± 0.84	74.12 ± 0.95	59.52 ± 0.77	58.62 ± 1.36	58.67 ± 0.85	$\mathbf{54.96 \pm 0.82}$
RPE-MulT	$\mathbf{76.35 \pm 0.63}$	$\mathbf{75.72 \pm 1.07}$	$\mathbf{75.99 \pm 0.58}$	$\mathbf{74.83 \pm 0.93}$	$\mathbf{61.03 \pm 0.75}$	$\mathbf{59.76 \pm 1.21}$	$\mathbf{59.98 \pm 0.90}$	54.17 ± 0.59

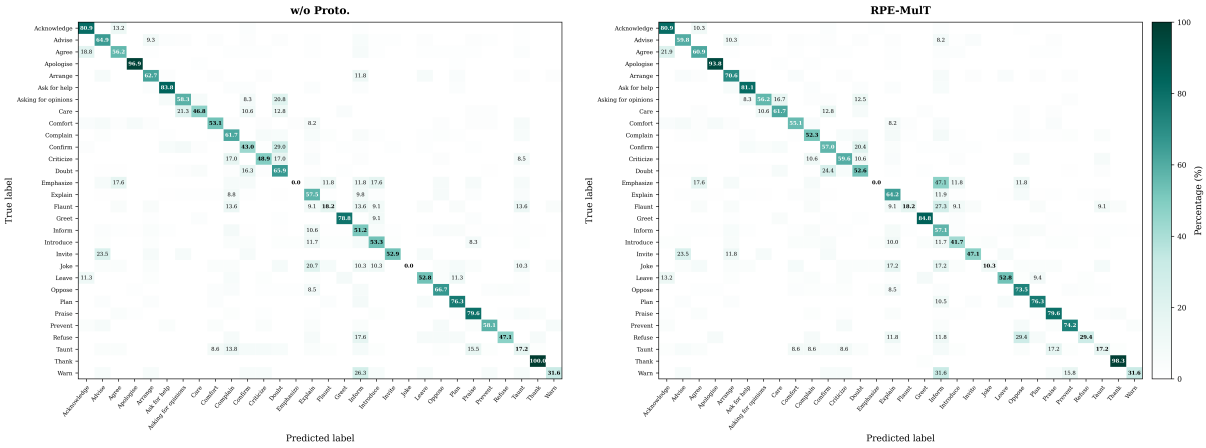

Figure 4: Row-normalized confusion matrices of w/o Proto. and RPE-MulT on MIntRec2.0. Darker diagonal cells indicate higher class-wise accuracy, while off-diagonal cells indicate misclassification patterns. The selected classes for prototype representation visualization, i.e., *Oppose*, *Explain*, *Doubt*, and *Advise*, correspond to representative high-confusion and semantically ambiguous intent categories.

classes have ambiguous decision boundaries and are suitable for evaluating inter-class separability.

Doubt is selected because it represents one of the most ambiguous uncertainty-related intents. It is frequently confused with semantically related categories such as *Confirm*. In particular, *Doubt* $\rightarrow$ *Confirm* is the most frequent confusion in RPE-MulT, with 33 cases, and remains highly frequent in w/o Proto., with 22 cases. Therefore, *Doubt* provides a useful case for examining whether prototype-enhanced learning can improve the compactness of difficult intent representations.

Advise is included to cover goal-oriented directive intents. It is consistently confused with semantically related directive categories such as *Arrange*, with 10 cases in RPE-MulT and 9 cases in w/o Proto. In addition, RPE-MulT also shows confusion from *Advise* to *Inform*. Including *Advise* prevents the visualization from focusing only on explanation- or uncertainty-related intents and improves the semantic diversity of the selected subset.

Overall, the selected classes cover representative high-confusion patterns, including bidirectional semantic confusion, uncertainty-related ambiguity, and goal-oriented directive confusion. Therefore, they provide a challenging and informative subset for analyzing whether the prototype-enhanced module improves class-level compactness and separability.

C.2 Visualization Settings

For prototype representation visualization, we use PCA and UMAP to project the high-dimensional fused utterance representations into a two-dimensional space. Specifically, we first concatenate the representations extracted from RPE-MulT and w/o Proto. to ensure that both models are projected under the same transformation. We then apply PCA implemented in scikit-learn (Pedregosa et al., 2011) to reduce the feature dimension to 50, with the random seed set to 42. After PCA preprocessing, we apply UMAP implemented by the `umap-learn` package for two-dimensional visualization. The UMAP parameters are set as follows: `n_neighbors=10`, `min_dist=0.5`, `spread=1.0`, and `random_state=42`. Other parameters are kept as the default settings of the package.

PCA preprocessing is used to reduce noise and computational cost before applying UMAP. UMAP is only used for qualitative visualization; all quan-

Table 11: Parameter settings for PCA and UMAP visualization.

Package / Method	Parameter	Value
scikit-learn PCA	`n_components`	50
scikit-learn PCA	`random_state`	42
UMAP	`n_neighbors`	10
UMAP	`min_dist`	0.5
UMAP	`spread`	1.0
UMAP	`random_state`	42

titative separability metrics are computed on the original high-dimensional fused representations before UMAP projection.

C.3 Representation Separability Metrics

We provide the detailed definitions of the representation separability metrics used in Section 4.5. Let $\mathcal{D} = \{(h_i, y_i)\}_{i=1}^{N}$ denote the set of selected samples, where h_i is the high-dimensional fused utterance representation and y_i is its intent label. For each class k, we denote the sample set as $\mathcal{D}_k = \{h_i \mid y_i = k\}$ and its class center as

$$\mu_k = \frac{1}{|\mathcal{D}_k|} \sum_{h_i \in \mathcal{D}_k} h_i. \tag{27}$$

The intra-class distance measures the average distance between samples and their corresponding class centers:

$$D_{\text{intra}} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{D}_k|} \sum_{h_i \in \mathcal{D}_k} d(h_i, \mu_k), \tag{28}$$

where K is the number of selected classes and $d(\cdot, \cdot)$ denotes the distance function.

The inter-class distance measures the average distance between different class centers:

$$D_{\text{inter}} = \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} d(\mu_i, \mu_j). \tag{29}$$

The inter/intra distance ratio is defined as:

$$R_{\text{sep}} = \frac{D_{\text{inter}}}{D_{\text{intra}} + \epsilon}, \tag{30}$$

where ϵ is a small constant for numerical stability.

The silhouette score evaluates how well each sample fits within its assigned class compared with its nearest neighboring class. For each sample h_i, let a_i be the average distance from h_i to other samples in the same class, and b_i be the minimum average distance from h_i to samples in another class. The silhouette score is computed as:

$$S = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}. \tag{31}$$

The Davies–Bouldin Index (DBI) evaluates clustering quality by comparing within-class dispersion and between-class separation:

$$\text{DBI} = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)}, \tag{32}$$

where σ_i denotes the average distance between samples in class i and their class center μ_i. A lower DBI indicates better clustering quality.