FROM EVIDENCE CONTENT TO DECISION ROLES: ROLE-CONDITIONED REASONING FOR MULTIMODAL INTENT RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal intent recognition (MIR) integrates linguistic, acoustic, and visual evidence for inferring a speaker's intent. Existing methods primarily model what this evidence conveys or how strongly it should contribute to fusion, leaving a more fundamental question underexplored: what role should the evidence play in the current decision? Reliable evidence may be redundant, while conflicting evidence can either correct predictions or introduce noise. Therefore, we propose **RCER**, a **R**ole-**C**onditioned **E**vidence **R**easoning framework that organizes multimodal intent inference into an evidence–role–decision pipeline. RCER employs a latent evidence structuring mechanism, which uses shared latent queries to organize heterogeneous modality sequences into slot-aligned, modality-specific evidence and explicitly encodes pairwise relations within each slot. It then performs role-conditioned evidence reasoning to characterize sample-dependent evidence roles through predictive conflict and signed counterfactual utility learned from leave-one-modality-out supervision. RCER encodes these signals as explicit role tokens that condition joint reasoning over unimodal and relational evidence, thereby guiding evidence interactions beyond scalar fusion weighting. Extensive experiments on two challenging MIR benchmarks, MIntRec and MIntRec2.0, demonstrate that RCER achieves state-of-the-art performance, highlighting the value of modeling evidence roles alongside evidence content for MIR.

1 INTRODUCTION

Multimodal intent recognition (MIR) integrates textual, acoustic, and visual signals to infer fine-grained user intents from both verbal and non-verbal cues (Zhu et al., 2025). Recent MIR research has increasingly moved beyond direct multimodal fusion. One line of work improves the quality and alignment of multimodal features through prompting, contrastive learning, contextual modeling, redundancy suppression, semantic disentanglement, and enhanced non-verbal modeling (Zhou et al., 2024; Huang et al., 2024; Sun et al., 2024; Zhu et al., 2024; Chen et al., 2024; Gong et al., 2025). Another line explicitly models sample-dependent modality contribution, confidence, or uncertainty to adapt the fusion process (Zhao et al., 2024; Hu et al., 2025; Yang et al., 2025a; Gong et al., 2026; Yang et al., 2025b). More recently, reasoning-oriented approaches capture fine-grained semantic relations and multimodal inconsistency rather than relying solely on feature-level fusion (Zhou et al., 2025; Wang et al., 2026).

Here, we use the term *evidence* to denote a task-conditioned representation that a modality provides for the intent decision. Existing methods characterize several important properties of such evidence: *what* it conveys, *how reliable* it appears, and *how it relates* to other modalities. What remains less explicit is how it affects the current joint decision conditioned on the surrounding evidence. Reliable evidence may be correct yet redundant when the remaining modalities are already sufficient; similarly, conflicting evidence may either provide a useful correction or introduce harmful interference. Conversely, weak standalone evidence may still become useful through complementary interaction with other modalities. We refer to this sample-dependent conditional effect as the decision *role* of evidence—how each evidence source should participate in the current prediction given the surrounding multimodal context.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

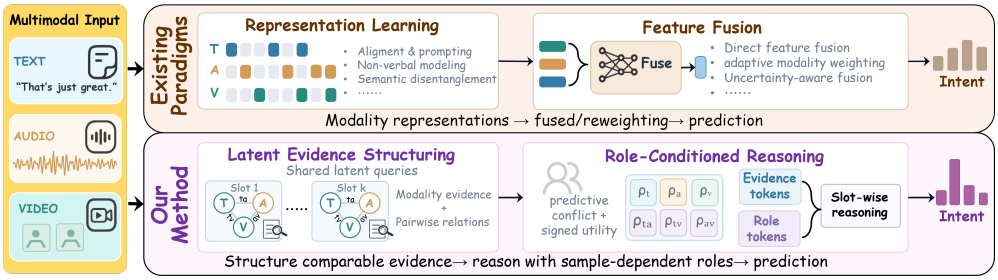


Figure 1: From evidence content to decision roles. Typical MIR paradigms characterize modalities primarily by their content or contribution, whereas RCER first structures multimodal evidence and then conditions evidence interactions on sample-dependent decision roles.

Taken together, these observations suggest that effective MIR requires reasoning beyond the properties of multimodal evidence itself. It must also determine how each evidence source should function within the current joint decision. This perspective introduces two challenges for role-aware MIR. First, heterogeneous modality streams must be organized into *comparable evidence units* before their relations and decision effects can be meaningfully assessed. Direct reasoning over raw modality representations is difficult because text, audio, and video differ substantially in temporal granularity and modality-specific structure; what is therefore required is structural comparability, rather than forcing their representations to be identical. Second, role characterization must capture the *direction* of an evidence source's marginal decision effect. Reliability indicates apparent trustworthiness and conflict describes cross-modal disagreement, but neither determines whether retaining an evidence source is beneficial, redundant, or harmful. Role-aware reasoning therefore requires both relational context and a signed measure of decision utility.

To address these challenges, we propose **RCER**, a **R**ole-**C**onditioned **E**vidence **R**easoning framework for MIR. RCER first performs **Latent Evidence Structuring**, where a shared bank of latent queries organizes heterogeneous modality streams into slot-aligned evidence under common latent coordinates and constructs explicit pairwise relational evidence. It then operationalizes sample-dependent decision roles using two complementary signals: predictive conflict captures the relational context of each modality, while signed counterfactual utility quantifies whether retaining its evidence benefits or harms the current prediction. Rather than collapsing these signals into scalar fusion weights, RCER maps them to explicit role tokens and jointly reasons over evidence content and role information within each latent slot before global decision aggregation. In this way, RCER separates *what evidence is available* from *how it should participate*, while allowing the two aspects to interact during reasoning. Experiments on MIntRec and MIntRec2.0 demonstrate the consistent effectiveness of this formulation. Our contributions are threefold:

- We formulate MIR from an **evidence–role–decision** view, distinguishing the information carried by task-conditioned evidence from its sample-dependent role in the current joint decision.
- We develop **Latent Evidence Structuring**, which uses shared latent evidence queries to organize heterogeneous modality streams into slot-aligned modality evidence and explicit pairwise relational evidence.
- We propose **Role-Conditioned Evidence Reasoning**, which characterizes evidence roles via predictive conflict and signed counterfactual utility, and represents them as explicit role tokens to condition slot-wise evidence reasoning.

2 Related Work

Multimodal Intent Recognition. MIR infers fine-grained intentions by exploiting linguistic, acoustic, and visual cues. MIntRec established a unified tri-modal benchmark, and MIntRec2.0 further expanded its data scale and intent taxonomy with out-of-scope intent detection (Zhang et al., 2022; 2024). Building on these benchmarks, recent MIR methods improve multimodal understanding from several complementary directions. For instance, modality-aware prompting and contrastive learning strengthen the alignment between textual and non-verbal cues (Zhou et al., 2024), while

progressive cross-modal interaction enables heterogeneous modalities to be aligned and fused in a shallow-to-deep manner (Huang et al., 2024). Contextual modeling further reduces intent ambiguity by exploiting cross-video information (Sun et al., 2024), and enhanced non-verbal representation improves the modeling of acoustic and visual cues (Gong et al., 2025). In addition, semantic synchronization aligns multimodal representations with semantic descriptions to promote more consistent multimodal understanding (Shen et al., 2025). These advances primarily strengthen the representation and interaction of multimodal cues, providing informative inputs for intent prediction.

Structured Multimodal Evidence Representation. Beyond direct feature fusion, a growing body of research explicitly organizes heterogeneous multimodal information to preserve task-relevant structure. InMu-Net suppresses intent-irrelevant redundancy through an information bottleneck (Zhu et al., 2024), while DuoDN disentangles common semantics from modality-specific information (Chen et al., 2024). SeD-UD further employs hierarchical information decoupling to separate redundancy and noise (Li et al., 2026), and related multimodal approaches decompose representations into shared and modality-private components (Meng et al., 2026) or model uncertainty at a finer semantic level (Zhu et al., 2026a). These studies demonstrate the value of structuring multimodal information rather than treating each modality as a monolithic feature vector. Their primary focus, however, remains on *what information to represent, preserve, or separate*. RCER builds on this perspective by organizing heterogeneous streams into comparable latent and relational evidence, which subsequently serves as the unit for decision-role reasoning.

Evidence Valuation and Reasoning. Another line of work focuses on how multimodal information should affect the prediction. Adaptive approaches estimate sample-dependent modality importance, reliability, or confidence to regulate fusion, as in MVCL-DAF (Hu et al., 2025), ARL (Yang et al., 2025a), DUMIR (Gong et al., 2026), and confidence-aware expert routing (Liu et al., 2026; Zhu et al., 2026b). More recent methods move beyond independent modality valuation toward explicit reasoning over cross-modal dependencies. LGSRR models fine-grained semantic importance, complementarity, and inconsistency through relational reasoning (Zhou et al., 2025), while CDPR separates consensus- and inconsistency-oriented pathways and adapts them according to multimodal conflict (Wang et al., 2026). Related causal and counterfactual approaches further estimate the marginal effect of modality-specific information on downstream predictions (Wang et al., 2025).

3 METHOD

3.1 TASK FORMULATION AND OVERVIEW

Task Formulation. Let $\mathcal{M} = \{t, a, v\}$ denote the textual, acoustic, and visual modalities. Given a multimodal sample $\mathcal{X} = \{x_m\}_{m\in\mathcal{M}}$ and its intent label $y \in \{1, \dots, C\}$, where C is the number of intent classes, MIR aims to learn a predictor $f_{\boldsymbol{\theta}}$ that maps $\mathcal{X}$ to intent logits $\mathbf{o} = f_{\boldsymbol{\theta}}(\mathcal{X}) \in \mathbb{R}^C$. The predicted intent is $\hat{y} = \arg\max_{c\in\{1,\dots,C\}} o_c$.

Framework Overview. As shown in Figure 2, RCER follows an evidence–role–decision pipeline. First, *Latent Evidence Structuring* maps heterogeneous modality sequences into a shared set of latent evidence slots. A shared query bank extracts modality-specific evidence under common latent coordinates, while an explicit relation encoder constructs pairwise cross-modal evidence within each slot. Second, *Role-Conditioned Evidence Reasoning* characterizes the sample-dependent state of this evidence using predictive conflict and signed counterfactual utility. These signals are lifted into explicit role tokens and jointly reasoned with the structured evidence. Finally, *Decision Aggregation* fuses the resulting role-conditioned slot representations into the final prediction.

3.2 LATENT EVIDENCE STRUCTURING

For each modality $m \in \mathcal{M}$, an encoder produces a temporal representation $\mathbf{F}_m \in \mathbb{R}^{L_m\times d_m}$. A modality-specific one-dimensional convolution projects the heterogeneous features into a common d-dimensional space: $\mathbf{H}_m = \mathrm{Conv1D}_m(\mathbf{F}_m; k_m) \in \mathbb{R}^{L_m\times d}$, where k_m denotes the convolution kernel size. Backbone choices and input features are detailed in Appendix B.

Shared-Query Evidence Extraction. Directly aligning heterogeneous temporal sequences tends to entangle modality-specific variations with decision-relevant correspondences. We therefore intro-

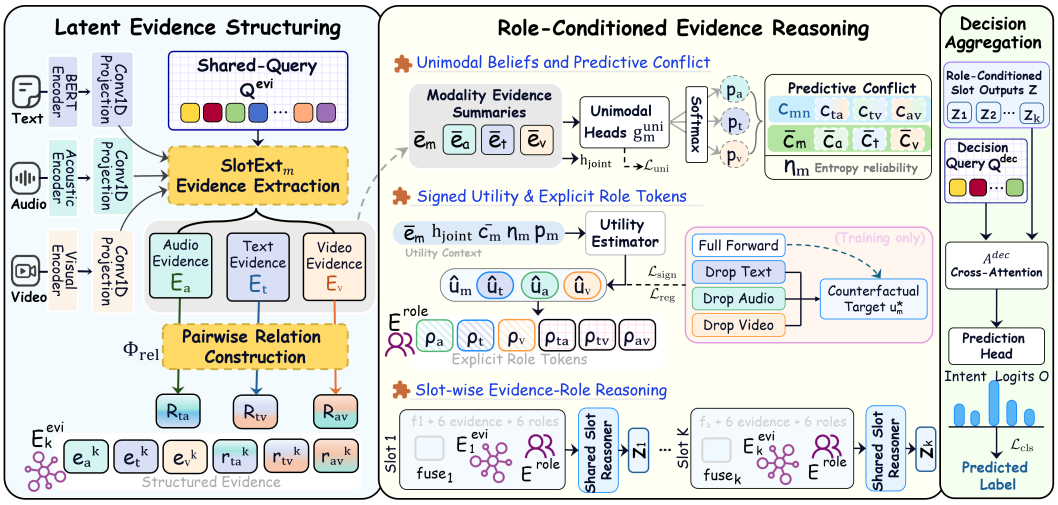


Figure 2: Overview of RCER. Heterogeneous modality streams are organized into slot-aligned modality evidence and pairwise relational evidence. Predictive conflict and signed counterfactual utility then characterize the sample-dependent role of this evidence, which conditions slot-wise reasoning before global decision aggregation.

duce a shared bank of k learnable evidence queries $\mathbf{Q}^{\mathrm{evd}} = [\mathbf{q}_1; \ldots; \mathbf{q}_k] \in \mathbb{R}^{k \times d}$, which provides a common set of latent coordinates across modalities. For each modality, the same query bank attends to its full temporal sequence through a modality-specific cross-attention extractor:

$$[\mathbf{e}_m^1; \ldots; \mathbf{e}_m^k] = \mathrm{SlotExt}_m(\mathbf{Q}^{\mathrm{evd}}, \mathbf{H}_m) \in \mathbb{R}^{k \times d}, \quad m \in \mathcal{M} \tag{1}$$

Here, $\mathrm{SlotExt}_m$ consists of multi-head cross-attention followed by residual feed-forward refinement and layer normalization. The shared query bank $\mathbf{Q}^{\mathrm{evd}}$ establishes a common latent coordinate across modalities, while modality-specific extractors preserve source-dependent information, yielding structurally comparable evidence without enforcing explicit cross-modal matching. Accordingly, $\mathbf{e}_m^k$ denotes the evidence contributed by modality m under the k-th latent coordinate.

Pairwise Relation Construction. The modality evidence vectors $\mathbf{e}_m^k$ describe modality-specific content but do not explicitly encode how two modalities interact within the same slot. For each modality pair $(m, n) \in \mathcal{P} = \{(t, a), (t, v), (a, v)\}$ and slot k, we construct an evidence relation:

$$\mathbf{s}_{mn}^k = [\mathbf{e}_m^k; \mathbf{e}_n^k; |\mathbf{e}_m^k - \mathbf{e}_n^k|; \mathbf{e}_m^k \odot \mathbf{e}_n^k] \in \mathbb{R}^{4d}, \quad \mathbf{r}_{mn}^k = \Phi_{\mathrm{rel}}(\mathbf{s}_{mn}^k) + \mathbf{i}_{mn}^{\mathrm{rel}} \in \mathbb{R}^d, \tag{2}$$

where Φ_{rel} is an MLP shared across modality pairs, $\mathbf{i}_{mn}^{\mathrm{rel}}$ is a learnable pair-identity embedding, and $\odot$ denotes element-wise multiplication. The descriptor preserves the content of both evidence endpoints while explicitly exposing their discrepancy and element-wise agreement. Together, the modality and relational evidence form a complete pairwise evidence structure within each latent slot k, which jointly encodes modality-specific content and pairwise cross-modal relations.

$$\mathbf{E}_k^{evi} = \{\mathbf{e}_m^k\}_{m \in \mathcal{M}} \cup \{\mathbf{r}_{mn}^k\}_{(m,n) \in \mathcal{P}} \tag{3}$$

3.3 Role-Conditioned Evidence Reasoning

The structured evidence specifies what information is available and how modality contents relate within each latent slot, but it does not determine whether a source is beneficial, redundant, or interfering in the current decision. We therefore characterize evidence role through two complementary decision-level signals: predictive conflict describes the relational context surrounding an evidence source, while signed counterfactual utility captures the direction of its marginal decision effect.

Unimodal Beliefs and Predictive Conflict. To characterize how evidence behaves at the decision level, we first summarize the latent evidence of each modality $\bar{\mathbf{e}}_m \in \mathbb{R}^d$ and derive a unimodal intent belief. For $m \in \mathcal{M}$, its K evidence vectors are mapped to a unimodal distribution $\mathbf{p}_m \in \mathbb{R}^C$:

$$\bar{\mathbf{e}}_m = \frac{1}{K} \sum_{k=1}^{K} \mathbf{e}_m^k, \qquad \mathbf{p}_m = \mathrm{softmax}\left(g_m^{\mathrm{uni}}(\bar{\mathbf{e}}_m)\right), \tag{4}$$

where $g_m^{\text{uni}} : \mathbb{R}^d \to \mathbb{R}^C$ is a lightweight modality-specific classification head. These heads are explicitly supervised during training so that $\mathbf{p}_m$ represents an intent-level belief rather than an arbitrary projection of the modality features. For $(m, n) \in \mathcal{P}$, we measure predictive disagreement c_{mn} using normalized Jensen–Shannon divergence:

$$c_{mn} = \text{JSD}_2(\mathbf{p}_m, \mathbf{p}_n) \in [0, 1]; \qquad \bar{c}_m = \frac{1}{|\mathcal{M}| - 1} \sum\nolimits_{n \in \mathcal{M} \setminus \{m\}} c_{mn}, \tag{5}$$

where $\text{JSD}_2(\mathbf{p}_m, \mathbf{p}_n) = \frac{1}{2}\,\text{KL}_2(\mathbf{p}_m \| \mathbf{q}_{mn}) + \frac{1}{2}\,\text{KL}_2(\mathbf{p}_n \| \mathbf{q}_{mn})$, $\mathbf{q}_{mn} = \frac{1}{2}(\mathbf{p}_m + \mathbf{p}_n)$. JSD is used because conflict is defined between predictive distributions and should not depend on modality ordering. With base-2 logarithms, it is symmetric and bounded in $[0, 1]$, providing a common scale for comparing disagreement across modality pairs and samples. Here, c_{mn} measures pairwise predictive conflict, while $\bar{c}_m$ summarizes the conflict context surrounding modality m.

Signed Counterfactual Utility. Conflict is non-directional: disagreement may be corrective or detrimental. To capture this distinction, we define the signed counterfactual utility of modality m as the change in ground-truth cross-entropy when its evidence is removed. To avoid circular role supervision, targets are generated by a *role-free* branch. Let $\mathbf{o}^{\text{full}}$ denote its logits using all modalities, and let $\mathbf{o}^{-m}$ denote the logits obtained after masking modality m together with its incident relational evidence. With $\ell_y(\mathbf{o}) = \text{CE}(\mathbf{o}, y)$, the training target is

$$u_m^\star = \tanh\big((\ell_y(\mathbf{o}^{-m}) - \ell_y(\mathbf{o}^{\text{full}}))/\tau_u\big) \in [-1, 1]. \tag{6}$$

A positive value indicates that removing the modality increases the loss, and the modality is therefore beneficial for that sample; a negative value indicates interference, while values near zero indicate limited marginal effect. This quantity is a supervised counterfactual utility signal rather than a causal-effect estimate. Because y and leave-one-modality-out inference are unavailable at test time, RCER learns a modality-specific utility estimator. We first form a global evidence context $\mathbf{h}_{\text{joint}} = g_{\text{joint}}([\bar{\mathbf{e}}_t; \bar{\mathbf{e}}_a; \bar{\mathbf{e}}_v])$, and predict $\hat{u}_m$.

$$[\widetilde{u}_m, s_m] = g_m^u([\bar{\mathbf{e}}_m; \mathbf{h}_{\text{joint}}; \mathbf{p}_m; \eta_m; \bar{c}_m]), \quad \hat{u}_m = \tanh(\widetilde{u}_m) \in [-1, 1], \tag{7}$$

Here, $\eta_m = 1 - \mathcal{H}(\mathbf{p}_m)/\log C$ is an entropy-based reliability proxy, with $\mathcal{H}(\cdot)$ Shannon entropy and s_m a sign logit. The evidence-derived inputs to the utility pathway are detached from the evidence backbone, so utility supervision trains the utility modules without directly reshaping the learned evidence through this auxiliary path. At inference, $\hat{u}_m$ is obtained in a single forward pass.

Explicit Role Tokenization. We encode the resulting decision states in the same d-dimensional space as the structured evidence. The role representation mirrors the topology of the evidence structure: each modality evidence source receives a modality-level role token, and each pairwise relation receives a relation-level role token. For modality m and modality pair $(m, n) \in \mathcal{P}$, we construct

$$\boldsymbol{\rho}_m = g_{\text{mod}}(\text{sg}([\bar{c}_m; \hat{u}_m])) + \mathbf{i}_m^{\text{role}}, \quad \boldsymbol{\rho}_{mn} = g_{\text{pair}}(\text{sg}([c_{mn}; \hat{u}_m; \hat{u}_n])) + \mathbf{i}_{mn}^{\text{role}}, \tag{8}$$

where g_{mod} and g_{pair} are role encoders shared across modalities and pairs, respectively, and $\mathbf{i}_m^{\text{role}}, \mathbf{i}_{mn}^{\text{role}} \in \mathbb{R}^d$ are learnable identity embeddings that preserve the source of each role token. The stop-gradient operator $\text{sg}(\cdot)$ prevents the final classification objective from changing the role signals through the tokenization path. This yields three modality-role tokens $\{\boldsymbol{\rho}_t, \boldsymbol{\rho}_a, \boldsymbol{\rho}_v\}$ and three relation-role tokens $\{\boldsymbol{\rho}_{ta}, \boldsymbol{\rho}_{tv}, \boldsymbol{\rho}_{av}\}$, forming an explicit sample-level role context.

Slot-wise Evidence–Role Reasoning. Evidence content is specific to each latent slot, whereas the estimated roles describe the sample-level decision state of their sources. The same role context is therefore shared across slots, while the slot reasoner determines how that context should affect different latent evidence coordinates. For slot k, we have

$$\mathbf{E}_k^{evi} = [\mathbf{e}_t^k; \mathbf{e}_a^k; \mathbf{e}_v^k; \mathbf{r}_{ta}^k; \mathbf{r}_{tv}^k; \mathbf{r}_{av}^k] \in \mathbb{R}^{6 \times d}, \quad \mathbf{E}^{role} = [\boldsymbol{\rho}_t; \boldsymbol{\rho}_a; \boldsymbol{\rho}_v; \boldsymbol{\rho}_{ta}; \boldsymbol{\rho}_{tv}; \boldsymbol{\rho}_{av}] \in \mathbb{R}^{6 \times d}. \tag{9}$$

With a learnable fusion token $fuse_k \in \mathbb{R}^d$, the slot input is $\mathbf{X}_k = [fuse_k; \mathbf{E}_k^{\text{evi}}; \mathbf{E}^{\text{role}}] \in \mathbb{R}^{13 \times d}$. A shared reasoner performs self-attention jointly over evidence content and role context:

$$\widetilde{\mathbf{X}}_k = \mathcal{R}_{\text{slot}}(\mathbf{X}_k), \qquad \mathbf{z}_k = [\widetilde{\mathbf{X}}_k]_1 \in \mathbb{R}^d, \quad k = 1, \ldots, K, \tag{10}$$

where $\mathcal{R}_{\text{slot}}$ is a shared Transformer-based reasoning module and $[\cdot]_1$ selects the output corresponding to the fusion token. Thus, $\mathbf{z}_k$ summarizes the k-th evidence slot after its content interactions have been conditioned on the current role state. Consequently, role information influences *how evidence interacts* within a slot instead of acting only as a scalar rescaling of evidence magnitude.

3.4 Decision Aggregation and Learning

Decision Query Aggregation. After role-conditioned reasoning, RCER obtains $\mathbf{Z} = [\mathbf{z}_1; \ldots; \mathbf{z}_K]$, where each $\mathbf{z}_k$ summarizes the evidence in one role-conditioned latent slot. Because different slots may capture complementary intent cues, uniform pooling can obscure their heterogeneous contributions. We therefore use M learnable decision queries $\mathbf{Q}^{\mathrm{dec}} \in \mathbb{R}^{M \times d}$ to retrieve task-relevant information from the complete slot set:

$$\mathbf{D} = \mathcal{A}_{\mathrm{dec}}(\mathbf{Q}^{\mathrm{dec}}, \mathbf{Z}) \in \mathbb{R}^{M \times d}, \quad \widetilde{\mathbf{h}} = \mathrm{vec}(\mathbf{D}) + g_{\mathrm{pred}}(\mathrm{vec}(\mathbf{D})), \quad \mathbf{o} = \mathbf{W}_o \widetilde{\mathbf{h}} + \mathbf{b}_o, \tag{11}$$

where $\mathcal{A}_{\mathrm{dec}}$ is a residual cross-attention block with feed-forward refinement, g_{pred} is a lightweight residual prediction network, and $\mathbf{W}_o, \mathbf{b}_o$ are the final classifier parameters.

Learning Objectives. Training jointly supervises the final intent prediction, the unimodal belief heads used for conflict estimation, and the utility estimator. The primary classification loss and unimodal auxiliary loss are:

$$\mathcal{L}_{\mathrm{cls}} = \mathrm{CE}(\mathbf{o}, y), \quad \mathcal{L}_{\mathrm{uni}} = \frac{1}{|\mathcal{M}|} \sum\nolimits_{m \in \mathcal{M}} \mathrm{CE}(g_m^{\mathrm{uni}}(\bar{\mathbf{e}}_m), y). \tag{12}$$

The unimodal supervision anchors $\mathbf{p}_m$ to intent semantics, making the predictive disagreement in equation 5 meaningful at the decision level. For utility learning, we regress the counterfactual target magnitude and explicitly supervise its polarity:

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{|\mathcal{M}|} \sum\nolimits_{m \in \mathcal{M}} \mathrm{SmoothL1}(\hat{u}_m, u_m^\star), \quad \mathcal{L}_{\mathrm{sign}} = \frac{1}{|\mathcal{M}|} \sum\nolimits_{m \in \mathcal{M}} \mathrm{BCE}_{\mathrm{bal}}(s_m, s_m^\star), \tag{13}$$

where $s_m^\star = \mathbf{1}[u_m^\star > 0]$, and $\mathrm{BCE}_{\mathrm{bal}}$ denotes class-balanced binary cross-entropy over positive and negative utility signs. The complete objective is $\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{uni}} \mathcal{L}_{\mathrm{uni}} + \lambda_u \left(\mathcal{L}_{\mathrm{reg}} + \lambda_s \mathcal{L}_{\mathrm{sign}} \right)$, where λ_{uni}, λ_u, and λ_s balance the corresponding auxiliary supervision terms.

Progressive Training and Inference. The training schedule follows the dependency of the proposed reasoning chain: utility estimation requires meaningful structured evidence and unimodal beliefs, while role-conditioned reasoning in turn depends on a sufficiently trained utility estimator. We therefore first optimize the evidence backbone and unimodal heads, then activate counterfactual utility supervision, and finally enable explicit role conditioning and optimize the complete model. Counterfactual targets are generated without gradient propagation through the target-generation path, while the stop-gradient operation in equation 8 prevents the final classification objective from directly modifying the role signals. At inference, the counterfactual target-generation path is removed; RCER estimates $\hat{u}_m$ directly and performs intent prediction in a single forward pass.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate RCER on two challenging benchmark datasets for MIR: (i) MIntRec, a fine-grained MIR benchmark comprising 2,224 annotated samples with textual, visual, and acoustic modalities across 20 intent categories; and (ii) MIntRec2.0, a larger extension of MIntRec featuring richer conversational contexts and 30 intent categories.

Baselines. We compare RCER against a comprehensive set of state-of-the-art MIR methods, including MISA (Hazarika et al., 2020), MAG-BERT (Rahman et al., 2020), MulT (Tsai et al., 2019), TCL-MAP (Zhou et al., 2024), MVCL-DAF (Hu et al., 2025), and DUMIR (Gong et al., 2026). Detailed descriptions are provided in Appendix A.

Evaluation Protocol. We evaluate performance using Accuracy (ACC), macro-averaged Recall (R), weighted F1 (WF1), and weighted Precision (WP). Higher values indicate better performance for all metrics. All baselines are reproduced under the same official data splits, random seeds, and metric implementation, with method-specific settings following their official implementations. All reported results are averaged over three random seeds. Appendix B provides full experimental details. Appendix C further reports sensitivity analyses of the numbers of latent evidence slots k and decision slots M, the utility supervision weight λ_u, and the counterfactual utility temperature τ_u.

Table 1: Comparison results on MIntRec and MIntRec2.0. **Bold** values denote the best results, and underlined values denote the second best results.

Method	MIntRec				MIntRec2.0			
	ACC(%)	R(%)	WF1(%)	WP(%)	ACC(%)	R(%)	WF1(%)	WP(%)
MAG-BERT	70.86	67.34	70.55	71.61	59.91	52.57	55.67	52.45
MISA	71.46	68.72	71.93	73.57	57.27	51.31	57.21	57.63
MulT	71.09	67.39	71.05	71.83	59.89	51.58	53.53	51.79
TCL-MAP	73.48	70.26	73.04	73.33	57.73	53.36	57.45	57.97
DUMIR	74.31	72.38	74.03	75.10	60.30	**53.93**	59.25	59.43
MVCL-DAF	75.13	73.39	75.11	75.53	59.79	52.94	58.63	58.98
RCER (Ours)	**77.16**	**73.58**	**76.88**	**77.43**	**60.96**	53.53	**59.61**	**60.42**

4.2 Main Comparisons

Table 1 compares RCER with representative MIR methods on MIntRec and MIntRec2.0. On MIntRec, RCER obtains 77.16% ACC, 73.58% recall, 76.88% WF1, and 77.43% WP, outperforming all competing methods on every reported metric. Compared with MVCL-DAF, the strongest baseline on this benchmark, RCER improves ACC, recall, WF1, and WP by 2.03, 0.19, 1.77, and 1.90 percentage points, respectively. The advantage remains consistent on the larger MIntRec2.0 benchmark. RCER achieves the highest 60.96% ACC, 59.61% WF1, and 60.42% WP, while attaining the second-best macro recall of 53.53%, only 0.40 percentage points below DUMIR. Since macro recall weights all intent classes equally, this small gap suggests a slightly different class-wise recall trade-off, whereas RCER provides stronger overall and support-weighted recognition performance. Overall, these results support the effectiveness of structuring multimodal observations as evidence and reasoning over their sample-dependent decision roles across both benchmarks.

4.3 Ablation Studies

We conduct ablation studies on MIntRec to examine three design questions underlying RCER: (i) how heterogeneous multimodal observations should be structured as evidence, (ii) how sample-dependent role information should be incorporated into reasoning, and (iii) whether the supervision used to construct these role signals is necessary. Table 2 summarizes the results. Each block varies one design axis while keeping the remaining components fixed. For the supervision ablation, a dynamic role signal is activated only when its corresponding supervision is available, yielding nested and well-defined training configurations.

Effect of Evidence Structuring. Settings (a)–(d) examine how multimodal observations should be organized before role-conditioned reasoning. Applying pairwise relations directly to pooled raw features provides little benefit over raw features alone, with WF1 changing from 73.66 to 73.72. In contrast, replacing raw pooling with shared-query latent extraction improves ACC from 73.93 to 75.13 and WF1 from 73.66 to 74.77. Combining latent evidence with relational evidence in the full model further raises these scores to 77.16 and 76.88, respectively. The matched-capacity control in (d) remains substantially below the full model, suggesting that the gain is not explained by additional parameterization alone. Together, these results indicate that relational modeling is most effective after heterogeneous observations have been organized into a shared, slot-aligned evidence space.

Effect of Role-Conditioned Reasoning. Settings (e), (f), and (i) use the same structured evidence but differ in how sample-dependent role information participates in prediction. Direct readout without explicit role conditioning reaches 76.03 ACC and 75.77 WF1, while compressing conflict and utility into scalar modality gates reduces performance to 75.21 ACC and 75.05 WF1. Explicit role tokens instead achieve 77.16 ACC and 76.88 WF1. Importantly, scalar gating and role tokens use the same conflict and utility signals; the difference lies in how these signals are represented and consumed. The result supports modeling evidence roles as explicit contextual representations that can interact with evidence content during reasoning, rather than reducing them to scalar weights.

Effect of Role Supervision. Settings (g)–(i) progressively introduce the supervision for constructing the role context. From classification-only training, modality-wise belief supervision improves ACC from 73.93 to 76.48 and WF1 from 73.70 to 76.24, providing supervised unimodal beliefs

Table 2: Ablation study results of the key components of RCER on MIntRec.

Ablation	Setting	ACC(%)	R(%)	WF1(%)	WP(%)
Evidence Structuring	(a) Raw features	73.93	71.10	73.66	74.51
	(b) Raw + pairwise relations	74.01	70.95	73.72	74.46
	(c) Latent evidence	75.13	71.79	74.77	75.19
	(d) Raw + matched capacity	73.82	70.44	73.86	74.48
Role-Conditioned Reasoning	(e) Direct readout	76.03	72.37	75.77	76.22
	(f) Scalar role gating	75.21	71.75	75.05	75.41
Role Supervision	(g) $\mathcal{L}_{\text{cls}}$ only	73.93	71.46	73.70	74.64
	(h) $+\mathcal{L}_{\text{uni}}$ (conflict)	76.48	73.14	76.24	76.88
Full Model	(i) $+\mathcal{L}_u$ (conflict + utility; RCER)	**77.16**	**73.58**	**76.88**	**77.43**

from which predictive conflict is derived. Adding signed counterfactual utility supervision further increases ACC and WF1 to 77.16 and 76.88, respectively. Thus, predictive disagreement alone does not fully characterize the role of modality evidence: signed utility contributes complementary directional information about whether retaining a modality benefits or harms the current decision. Overall, the three ablations support the central RCER design: heterogeneous observations are first organized into structured evidence, sample-dependent roles are represented explicitly during reasoning, and these roles are learned from complementary conflict and signed-utility supervision.

4.4 Analysis of Evidence Roles

Role-Signal Landscapes. We first examine the relationship among predictive reliability η_m, conflict context $\bar{c}_m$, and signed counterfactual utility $u_m^\star$. As shown in Figure 3, evidence with similar reliability and conflict can exhibit markedly different utilities, indicating that neither signal alone determines its directional effect on the final decision. The landscapes also reveal a clear modality asymmetry: textual evidence is concentrated in regions of high reliability and positive utility, whereas acoustic and visual evidence is substantially more dispersed and frequently has weak marginal utility. Nevertheless, low-reliability non-verbal evidence can still exhibit positive utility in context, suggesting that the challenge is not simply to suppress weak modalities, but to determine when their evidence becomes useful.

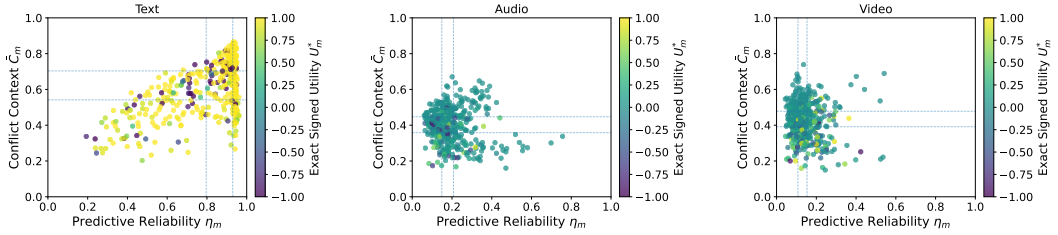


Figure 3: Evidence-role landscapes across modalities on MIntRec.

Representative Role Cases. Figure 4 provides four instance-level examples that make these distributional patterns concrete. **Case A** shows supportive text: the textual evidence is highly reliable, predicts the correct *Complain* intent, and has strongly positive utility. **Case B** shows redundancy: although audio correctly predicts *Praise*, its utility is nearly zero ($u_a^\star = 0.006$), indicating little marginal contribution beyond the remaining evidence. **Case C** shows misleading visual evidence: video favors the incorrect *Complain* intent and has negative utility ($u_v^\star = -0.041$), while RCER still recovers *Inform*. Conversely, **Case D** shows complementary visual evidence: despite low reliability ($\eta_v = 0.094$) and an incorrect unimodal prediction, the visual cue has positive utility ($u_v^\star = 0.257$) and contributes to the correct *Joke* decision. Together, the landscape and case analyses show that evidence role is sample-dependent and cannot be reduced to unimodal reliability, correctness, or disagreement alone. This provides both distribution-level and instance-level support for the evidence–role view underlying RCER.

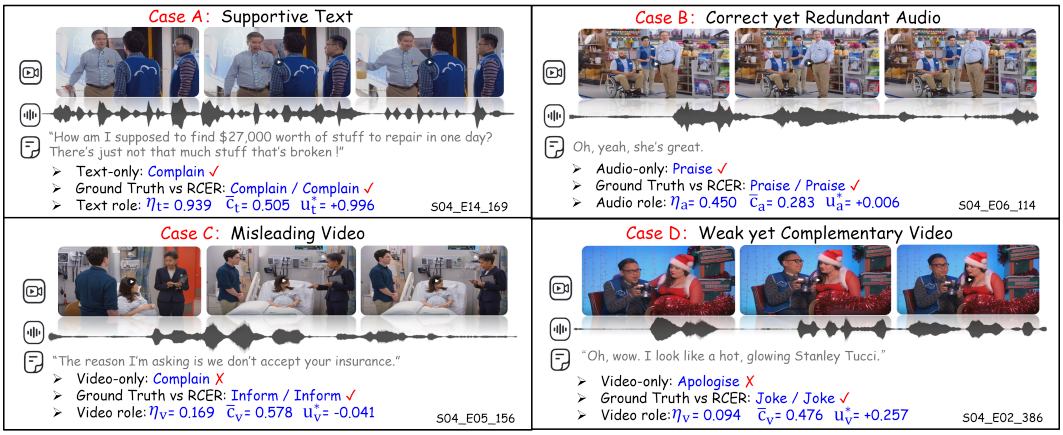


Figure 4: **Representative sample-specific evidence roles on MIntRec,** illustrating supportive, redundant, misleading, and complementary evidence under different multimodal contexts.

4.5 Robustness under Textual Evidence Corruption

We further evaluate the robustness of RCER when textual evidence is progressively corrupted. Following the same protocol for all methods, a fraction p of textual feature entries is replaced by the corresponding entries from another sample in the same mini-batch: $\mathbf{H}_t^\star = (1-\mathbf{P})\odot\mathbf{H}_t + \mathbf{P}\odot\mathbf{H}_t^\pi$, $P_i \sim \mathrm{Bernoulli}(p)$, where π denotes a random within-batch permutation and p controls the corruption severity, while the acoustic and visual inputs remain unchanged. As shown in Table 3, performance decreases substantially for all methods as the corruption rate increases. RCER nevertheless achieves the highest WF1 at every reported corruption level. The advantage is particularly pronounced under severe corruption: at $p = 0.7$ and $p = 0.9$, RCER exceeds the strongest competing method by 2.12 and 3.39 percentage points, respectively. These results indicate that RCER retains stronger predictive performance when textual evidence becomes increasingly unreliable, complementing its gains under clean evaluation. Complete results for all evaluation metrics are provided in Appendix D.

Table 3: Robustness to corrupted textual evidence on MIntRec (WF1, %).

p	MAG-BERT	MISA	MulT	TCL-MAP	DUMIR	MVCL-DAF	**RCER**
p=0.0	70.55	71.93	71.05	73.04	74.03	<u>75.11</u>	**76.88**
p=0.3	69.37	69.98	68.91	70.40	74.04	<u>74.23</u>	**75.36**
p=0.5	45.25	47.86	47.37	46.89	51.36	<u>52.02</u>	**53.46**
p=0.7	16.47	19.77	21.45	17.53	<u>24.43</u>	18.37	**26.55**
p=0.9	15.20	16.19	<u>16.63</u>	15.44	16.57	14.68	**20.02**

5 Conclusion

In this work, we revisit multimodal intent recognition from an ***evidence–role*** perspective: effective multimodal reasoning requires not only representing what each modality conveys, but also determining how its evidence should participate in the current decision. To operationalize this view, we propose **RCER**, which organizes heterogeneous modality streams into slot-aligned latent and relational evidence, characterizes sample-dependent roles through predictive conflict and signed counterfactual utility, and incorporates these roles as explicit representations for evidence-conditioned reasoning rather than scalar fusion weights. Across MIntRec and MIntRec2.0, RCER achieves the strongest performance among the compared methods. Controlled ablations further support the complementary contributions of structured evidence, explicit role representations, and role supervision, while evidence-role analyses show that reliability, disagreement, and decision utility capture distinct aspects of multimodal evidence behavior. Together, these results support evidence role as a complementary dimension to multimodal representation learning, providing a more expressive framework for sample-dependent multimodal intent reasoning.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, terminology consistency, LaTeX formatting, and the refinement of figures and manuscript presentation. They did not contribute to the creation of experimental data, the analysis of measurements, or the formulation of the paper's scientific conclusions. All content produced with AI assistance was examined by the authors, who remain fully responsible for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce the proposed method and the reported experiments throughout the paper and appendices. Section 3 specifies the RCER architecture, role construction, learning objectives, progressive training procedure, and inference process, while Section 4.1 describes the datasets, evaluation metrics, and comparison protocol. Appendix B further documents the input representations, dataset-specific model configurations, optimization settings, training schedule, and baseline reproduction protocol. All main benchmark comparisons are reproduced using the official data splits, the same random seeds and metric implementation, and the benchmark-provided multimodal inputs, with method-specific configurations following their official implementations whenever available.

REFERENCES

Zhanpeng Chen, Zhihong Zhu, Xianwei Zhuang, Zhiqi Huang, and Yuexian Zou. Dual-oriented disentangled network with counterfactual intervention for multimodal intent detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17554–17567. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.972.

Weiyin Gong, Kai Zhang, Yanghai Zhang, Qi Liu, Xinjie Sun, Junyu Lu, and Linbo Zhu. Wdmir: Wavelet-driven multimodal intent recognition. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 5226–5234, 2025. doi: 10.24963/ijcai.2025/582.

Weiyin Gong, Kai Zhang, Xinjie Sun, Yanghai Zhang, and Hui Li. DUMIR: Dynamic uncertainty-aware multimodal intent recognition. *Pattern Recognition*, 180:114474, 2026. doi: 10.1016/j.patcog.2026.114474.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020. doi: 10.1145/3394171.3413678.

Bo Hu, Kai Zhang, Yanghai Zhang, and Yuyang Ye. Adaptive multimodal fusion: Dynamic attention allocation for intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17267–17275, 2025. doi: 10.1609/aaai.v39i16.33898.

Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 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*, pp. 10206–10210, 2024. doi: 10.1109/ICASSP48485.2024.10446922.

Qin Li, Wenbo Zhang, Limei Liu, Han Peng, Junfeng Yang, and Guanying Xu. SeD-UD: An influence-driven and hierarchically-decoupled information bottleneck for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30346–30356, 2026.

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

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8791–8800, 2026.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369, 2020. doi: 10.18653/v1/2020.acl-main.214.

Yaomin Shen, Xiaojian Lin, and Wei Fan. A-mess: Anchor-based multimodal embedding with semantic synchronization for multimodal intent recognition. In *2025 IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209824.

Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. Contextual augmented global contrast for multimodal intent recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26953–26963, 2024. doi: 10.1109/CVPR52733.2024.02546.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656.

Jun Wang, Fuyuan Cao, Zhixin Xue, Xingwang Zhao, and Jiye Liang. CMoB: Modality valuation via causal effect for balanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2763.

Yifan Wang, Peiwu Wang, Yunxian Chi, Zhinan Gou, and Kai Gao. Mitigating multimodal inconsistency via cognitive dual-pathway reasoning for intent recognition. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 797–806. ACM, 2026. doi: 10.1145/3805622.3810808.

Qu Yang, Xiyang Li, Fu Lin, and Mang Ye. Adaptive re-calibration learning for balanced multimodal intention recognition. In *Advances in Neural Information Processing Systems*, volume 38, pp. 139496–139517, 2025a. doi: 10.52202/085713-4658.

Qu Yang, Qinghongya Shi, Tongxin Wang, and Mang Ye. Uncertain multimodal intention and emotion understanding in the wild. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24700–24709, 2025b. doi: 10.1109/CVPR52734.2025.02300.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. MIntRec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1688–1697, 2022. doi: 10.1145/3503161.3547906.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. 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*, pp. 46367–46393, 2024. URL https://openreview.net/forum?id=nY9nITZQjc.

Xiyuan Zhao, Huijun Li, Tianyuan Miao, Xianyi Zhu, Zhikai Wei, Lifen Tan, and Aiguo Song. Learning multimodal confidence for intention recognition in human-robot interaction. *IEEE Robotics and Automation Letters*, 9(9):7819–7826, 2024.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17114–17122, 2024. doi: 10.1609/aaai.v38i15.29656.

Qianrui Zhou, Hua Xu, Yifan Wang, Xinzhi Dong, and Hanlei Zhang. LLM-guided semantic relational reasoning for multimodal intent recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22221–22237. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1130.

Aoqiang Zhu, Min Hu, Yan Xing, and Yiming Tang. Decoupled sub-feature uncertainty modeling for robust multimodal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 6921–6931, 2026a.

Yitong Zhu, Yuxuan Jiang, Guanxuan Jiang, Bojing Hou, Peng Yuan Zhou, Ge Lin, and Yuyang Wang. QA-MoE: Towards a continuous reliability spectrum with quality-aware mixture of experts for robust multimodal sentiment analysis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31689–31703. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.1461.

Zhihong Zhu, Xuxin Cheng, Zhaorun Chen, Yuyan Chen, Yunyan Zhang, Xian Wu, Yefeng Zheng, and Bowen Xing. InMu-Net: Advancing multi-modal intent detection via information bottleneck and multi-sensory processing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 515–524, 2024. doi: 10.1145/3664647.3681623.

Zhihong Zhu, Fan Zhang, Yunyan Zhang, Jinghan Sun, Zhiqi Huang, Qingqing Long, Bowen Xing, and Xian Wu. A survey on multi-modal intent recognition: Recent advances and new frontiers. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15223–15236, 2025.

A Baseline Methods

We compare RCER with representative MIR methods that cover different fusion and representation learning paradigms.

MISA (Hazarika et al., 2020) decomposes multimodal representations into modality-invariant and modality-specific subspaces. It encourages the invariant representations to capture information shared across modalities while preserving modality-dependent characteristics in the private representations, which are subsequently combined for downstream prediction.

MAG-BERT (Rahman et al., 2020) introduces a Multimodal Adaptation Gate (MAG) to inject acoustic and visual information into the hidden representations of a pretrained BERT model. The gating mechanism adaptively controls the nonverbal shift applied to textual representations, providing a strong text-centered multimodal fusion baseline.

MulT (Tsai et al., 2019) models cross-modal interactions through directional cross-modal attention between pairs of modality sequences. By directly attending across temporally unaligned modalities, it captures long-range cross-modal dependencies before integrating the resulting representations for prediction.

TCL-MAP (Zhou et al., 2024) is specifically designed for multimodal intent recognition and combines Modality-Aware Prompting (MAP) with Token-Level Contrastive Learning (TCL). MAP incorporates acoustic and visual cues into the textual semantic space through modality alignment and cross-modal attention, while TCL exploits intent-label semantics to improve multimodal representation learning through token-level contrastive supervision.

MVCL-DAF (Hu et al., 2025) combines Dynamic Attention Allocation Fusion (DAF) with multi-view contrastive learning. DAF dynamically adjusts attention both within individual modalities and across modalities according to the input sample, while the multi-view contrastive objective encourages complementary representations from different modality views.

DUMIR (Gong et al., 2026) introduces an uncertainty-aware framework for MIR. It estimates sample-dependent modality reliability from semantic representations and second-order statistical variations, and subsequently uses the estimated uncertainty to adaptively calibrate modality contributions during multimodal fusion.

B Experimental Details

Input Representations. For both benchmarks, we use `bert-large-uncased` as the textual backbone and the pre-extracted acoustic and visual features provided by the corresponding datasets. On MIntRec, the maximum sequence lengths of the textual, visual, and acoustic streams are 30, 230, and 480, with input dimensions of 1,024, 1,024, and 768, respectively. On MIntRec2.0, the corresponding sequence lengths are 50, 180, and 400, with input dimensions of 1,024, 256, and 768. The three modality streams are projected into a common latent space using one-dimensional convolutions. We use projection kernels $(k_t, k_a, k_v) = (6, 1, 1)$ on MIntRec and $(5, 1, 1)$ on MIntRec2.0.

Model Configuration. The shared latent dimension is set to $d = 360$ on both benchmarks. RCER uses $K = 18$ latent evidence slots and $M = 6$ decision queries on MIntRec, whereas MIntRec2.0 uses $K = 6$ and $M = 4$. The latter configuration corresponds to the setting used for the final MIntRec2.0 results reported in the main paper. The slot-wise evidence–role reasoner contains two Transformer layers with four attention heads and an FFN expansion ratio of 2. The hidden dimension of the utility estimator and role encoders is 128. We apply a dropout rate of 0.4 to the textual representation before projection and a dropout rate of 0.2 in the final prediction refinement module.

Optimization and Training. RCER is optimized using Adam with learning rates of 3×10^{-5} on MIntRec and 1×10^{-6} on MIntRec2.0. The training batch size is 16, while validation and testing use a batch size of 8. Training proceeds for at most 100 epochs, with an early-stopping patience of 8 epochs and gradient values clipped to 0.5.

The loss weights are set to $\lambda_u = 0.2$ and $\lambda_s = 0.5$ on both benchmarks, with the counterfactual utility temperature fixed to $\tau_u = 1.0$. For unimodal belief supervision, we use $\lambda_{\text{uni}} = 0.1$ on MIntRec and $\lambda_{\text{uni}} = 0.15$ on MIntRec2.0. The latter setting is the one used to obtain the final MIntRec2.0 results. Label smoothing is set to 0.02 on MIntRec and disabled on MIntRec2.0.

Progressive Role Learning. We adopt a progressive training schedule so that dynamic role conditioning is introduced only after the underlying evidence representations and utility estimator have been sufficiently initialized. On MIntRec, counterfactual utility supervision is activated from epoch 5 and role-conditioned reasoning from epoch 9. On MIntRec2.0, the corresponding stages begin from epochs 6 and 11, respectively. Before role conditioning is enabled, the model learns the multimodal evidence representations and supervised unimodal beliefs without using dynamic role information in the final prediction process.

Once utility supervision becomes active, role-free full and leave-one-modality-out forward passes are used during training to construct signed counterfactual utility targets. These targets supervise the utility estimator through the regression and sign objectives. After the role-conditioning stage begins, the resulting predictive conflict and estimated signed utility are encoded into explicit role tokens and jointly reasoned with the structured evidence.

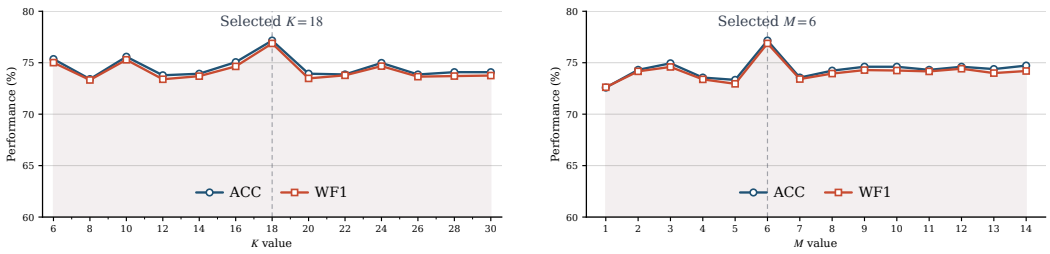
Model selection and early stopping for the role-enabled model are performed only after role conditioning has been activated. A `ReduceLROnPlateau` scheduler is used during this stage and reduces the learning rate by a factor of 0.1 when the validation loss ceases to improve. At inference time, the leave-one-modality-out branches used to construct counterfactual targets are removed. RCER therefore requires only a single role-enabled forward pass, with signed utility estimated directly by the learned utility estimator.

Baseline Reproduction Protocol. All baseline results in Table 1 are obtained from our own reproductions rather than copied from the corresponding publications. For every method, we use the official train/validation/test partitions and the same benchmark-provided textual, acoustic, and visual inputs used for RCER. All methods are evaluated using the same three random seeds and the same metric implementation. Method-specific architectural and optimization settings follow the corresponding official implementations or recommended configurations whenever available. This protocol controls variations in data partitioning, input features, random seeds, and evaluation procedures while preserving the method-specific design and training recipe of each baseline.

C Hyperparameter Sensitivity

We investigate four key hyperparameters on MIntRec. The numbers of latent evidence slots k and decision slots M determine the capacity for evidence organization and decision aggregation, whereas the utility supervision weight λ_u and counterfactual utility temperature τ_u control the learning of signed evidence utility. For each study, we vary one hyperparameter at a time while keeping all remaining settings fixed, and report ACC and WF1 results.

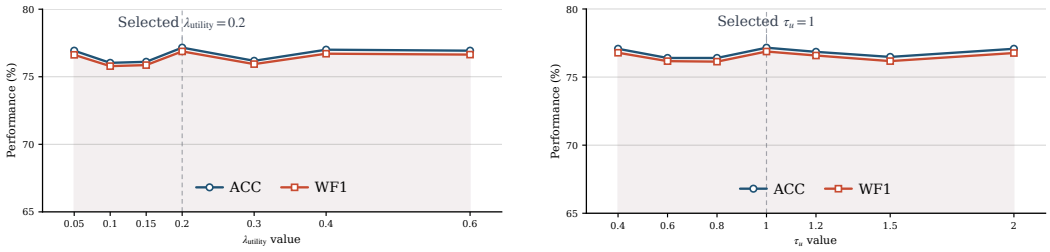
Architectural Sensitivity. Figure 5 examines the numbers of latent evidence slots k and decision slots M. Performance varies non-monotonically with both parameters, with the strongest ACC and WF1 obtained at $K = 18$ and $M = 6$. Using fewer slots restricts the capacity available for fine-grained evidence organization or decision aggregation, whereas further increasing the slot cardinality provides no consistent benefit. This behavior suggests that moderate slot cardinalities provide



(a) Latent evidence slots k (b) Decision slots M

Figure 5: **Sensitivity to architectural hyperparameters on MIntRec.** We vary the numbers of latent evidence slots k and decision slots M while keeping the remaining configuration fixed. Dashed lines indicate the settings adopted in RCER.

a favorable balance between representational capacity and redundant allocation. Accordingly, we adopt $K = 18$ and $M = 6$ on MIntRec.



(a) Utility loss weight λ_u (b) Utility temperature τ_u

Figure 6: **Sensitivity to role-learning hyperparameters on MIntRec.** We independently vary the utility supervision weight λ_u and counterfactual utility temperature τ_u while keeping all other settings fixed. Dashed lines indicate the adopted configuration, $\lambda_u = 0.2$ and $\tau_u = 1.0$.

Role-Learning Sensitivity. We next study the two hyperparameters directly associated with signed counterfactual utility learning. The coefficient λ_u controls the contribution of utility supervision relative to the primary classification objective, whereas τ_u controls the scale of the counterfactual utility target $\tanh(\Delta\ell/\tau_u)$.

As shown in Figure 6(a), RCER maintains competitive performance across a broad range of utility-supervision weights. When λ_u varies from 0.05 to 0.60, ACC ranges from 76.03% to 77.16%, while WF1 ranges from 75.79% to 76.88%. The adopted value $\lambda_u = 0.2$ achieves the strongest ACC and WF1 among the tested configurations. This result indicates that counterfactual utility supervision is effective without requiring an extreme weighting relative to the primary classification objective.

Figure 6(b) further examines the utility temperature. Across $\tau_u \in [0.4, 2.0]$, ACC varies from 76.40% to 77.16% and WF1 from 76.13% to 76.88%, with $\tau_u = 1.0$ again achieving the strongest result among the tested settings. The comparatively limited variation suggests that RCER does not rely on a narrowly tuned counterfactual target scale. Smaller temperatures produce more saturated signed utility targets, whereas larger temperatures compress utility differences toward zero; the selected intermediate value provides an effective balance between these behaviors.

Taken together, the two sensitivity studies reveal complementary behaviors. The architectural parameters exhibit clear preferred cardinalities, indicating that appropriate capacity is important for organizing and aggregating multimodal evidence. In contrast, role learning remains comparatively stable over reasonable ranges of λ_u and τ_u, with the adopted settings achieving the strongest performance among the tested configurations. These results support the selected RCER configuration without indicating dependence on extreme hyperparameter choices.

D COMPLETE ROBUSTNESS RESULTS

We provide the complete results of the textual corruption experiment described in Section 4.5. Table 4 reports Accuracy (ACC), macro-averaged Recall (R), weighted F1-score (WF1), and weighted Precision (WP) over all different corruption levels. The best and second-best results under each corruption level are highlighted in bold and underlined, respectively.

Table 4: Complete robustness results under progressive text corruption on MIntRec.

Rate	Method	ACC(%)	R(%)	WF1(%)	WP(%)
$p = 0.3$	MAG-BERT	70.79	66.64	69.37	70.60
	MISA	69.66	65.94	69.98	71.39
	MulT	69.66	65.36	68.91	69.11
	TCL-MAP	71.24	67.28	70.40	70.29
	DUMIR	<u>75.06</u>	<u>72.49</u>	74.04	<u>74.84</u>
	MVCL-DAF	74.61	71.92	<u>74.23</u>	74.44
	RCER	**75.96**	**73.35**	**75.36**	**76.37**
$p = 0.5$	MAG-BERT	47.87	41.32	45.25	45.10
	MISA	48.54	44.60	47.86	48.81
	MulT	49.44	45.91	47.37	48.00
	TCL-MAP	48.76	42.46	46.89	46.76
	DUMIR	<u>53.26</u>	48.71	51.36	<u>53.44</u>
	MVCL-DAF	52.81	<u>49.73</u>	<u>52.02</u>	51.92
	RCER	**54.61**	**50.40**	**53.46**	**56.31**
$p = 0.7$	MAG-BERT	16.63	13.24	16.47	16.64
	MISA	20.00	16.94	19.77	19.98
	MulT	21.80	18.33	21.45	21.43
	TCL-MAP	17.75	13.70	17.53	17.43
	DUMIR	<u>24.72</u>	<u>21.41</u>	<u>24.43</u>	<u>24.53</u>
	MVCL-DAF	18.65	16.12	18.37	18.28
	RCER	**26.97**	**23.80**	**26.55**	**26.54**
$p = 0.9$	MAG-BERT	15.28	12.08	15.20	15.45
	MISA	16.18	13.44	16.19	16.46
	MulT	<u>16.85</u>	12.64	<u>16.63</u>	16.55
	TCL-MAP	15.51	11.88	15.44	15.44
	DUMIR	16.63	<u>13.57</u>	16.57	<u>16.69</u>
	MVCL-DAF	14.83	12.39	14.68	14.60
	RCER	**20.00**	**17.72**	**20.02**	**20.21**