Conflict-Aware Trustworthy Representation Learning for Open-World Multimodal Intent Understanding

Anonymous ACL submission

Abstract

Open-world multimodal intent understanding frequently encounters cross-modal disagreement due to real-world noise or unseen semantics. However, existing methods typically assume modality consistency, which leads to entangled representations and degrades out-of-distribution (OOD) detection. We propose CARF, a trustworthy representation learning framework that quantifies cross-modal conflict via text-guided attention uncertainty and dynamically regulates modality contributions via conflict-aware gating. We also analyze the reshaped feature geometry. Experiments on MIntRec2.0, MELD-DA, and IEMOCAP-DA show CARF enhances open-world robustness while maintaining in-domain accuracy. Using Mahalanobis distance, CARF improves AUROC over additive fusion baselines from 68.87% to 71.93% on MELD-DA and from 89.71% to 94.86% on IEMOCAP-DA. Visualizations and geometric analyses confirm CARF forms compact in-domain clusters and a clear safety margin against OOD samples. Thus, explicitly modeling conflict crucially improves trustworthiness of multimodal representations under open-world uncertainty.

1 Introduction

Multimodal intent understanding aims to infer user intents from heterogeneous conversational signals such as language, facial expressions, and acoustic cues, and has become increasingly important in conversational AI systems (Zhang et al., 2021; Ouyang et al., 2021; Sun et al., 2024; Zhang et al., 2022, 2025a). Recent advances in pretrained language models and multimodal learning have substantially improved intent recognition performance through cross-modal interaction and feature fusion techniques (Zhang et al., 2025a). Existing studies mainly focus on learning richer multimodal representations by leveraging complementary semantic information across modalities (Sun et al., 2024; Zhang et al., 2022, 2025a; Tsai et al., 2019; Rahman et al., 2020).

However, real-world conversational environments are inherently open-world (Zhang et al., 2023, 2025a). Practical systems frequently encounter unseen intents, distribution shifts, environmental noise, and semantically inconsistent modality signals. Under such conditions, visual and acoustic modalities may become unreliable due to ambiguous facial expressions, corrupted audio, or unexpected interaction patterns. These factors introduce substantial uncertainty into multimodal representation learning and make reliable open-world intent understanding considerably more challenging.

Existing multimodal intent understanding methods mainly improve multimodal interaction through attention mechanisms, modality alignment, or representation fusion (Rahman et al., 2020; Tsai et al., 2019; Han et al., 2021; Zhou et al., 2024; Huang et al., 2024). Meanwhile, recent open-world intent detection approaches primarily focus on uncertainty estimation through confidence-based or distance-based scoring functions such as MSP, KNN, and Mahalanobis Distance (Hendrycks and Gimpel, 2017; Zhou et al., 2022; Podolskiy et al., 2021; Zhang et al., 2024a). Despite their effectiveness, most existing methods implicitly assume modality consistency and uniformly aggregate multimodal information, while paying limited attention to the reliability of multimodal representations themselves.

We argue that multimodal open-world uncertainty is fundamentally related to cross-modal conflict. Naively fusing unreliable modalities with semantic signals causes feature entanglement, which degrades in-distribution (ID) and out-of-distribution (OOD) separability, as conceptually illustrated in Figure 1(a). Therefore, modality conflict should not be discarded as optimization noise, but actively leveraged as a pivotal signal to learn

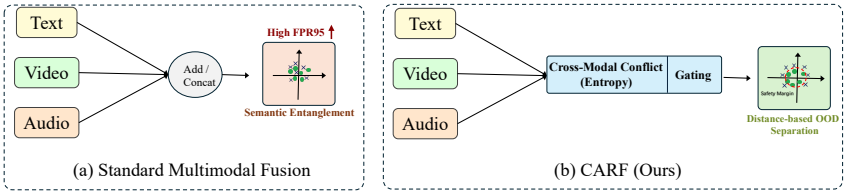

Figure 1: Conceptual comparison between standard fusion and our CARF. (a) Traditional methods blindly aggregate modalities, leading to semantic entanglement where OOD samples sneak into known ID clusters, rendering distance-based detection ineffective. (b) CARF explicitly estimates modality conflict via attention entropy to gate unreliable features, reshaping the geometry into a highly separable space with a clear safety margin.

reliability-aware representations.

To this end, we propose Conflict-Aware Representation Fusion (CARF). Instead of aggregating features, CARF quantifies modality conflict via text-guided cross-attention entropy. This conflict signal is dynamically utilized to suppress unreliable modalities through an amplified gating mechanism and adaptive noise regularization. By actively penalizing cross-modal disagreement, CARF structurally reshapes the feature space (Figure 1(b)). Experiments on MIntRec2.0 (Zhang et al., 2024a), MELD-DA (Saha et al., 2020), and IEMOCAP-DA (Saha et al., 2020) show that under a strict zero-OOD exposure setting, CARF maintains competitive ID accuracy while achieving state-of-the-art distance-based OOD detection, slashing FPR95 by 33.15% on IEMOCAP-DA (Saha et al., 2020). Geometric analyses confirm that CARF fundamentally purifies the representation manifold, providing robust safety boundaries for open-world environments.

Our contributions are summarized as follows:

- We present a paradigm shift in robust representation multimodal intent understanding by conceptualizing cross-modal conflict as a pivotal reliability signal rather than mere optimization noise under open-world uncertainty.
- We propose CARF, a framework that actively estimates epistemic uncertainty via cross-attention entropy, utilizing conflict-aware trust modulation and adaptive noise regularization to suppress unreliable modalities.
- Extensive experiments demonstrate that CARF yields highly separable feature geometry. Under a strict zero-OOD exposure setting, it establishes state-of-the-art distance-based OOD detection performance without degrading in-domain accuracy.

2 Related Work

2.1 Multimodal Intent Understanding

Multimodal intent understanding aims to infer user intents from heterogeneous conversational signals, including language, facial expressions, and acoustic cues (Sun et al., 2024; Zhang et al., 2022, 2025a, 2022, 2024a). Early multimodal methods mainly focused on feature-level fusion and cross-modal interaction modeling. MAG-BERT (Rahman et al., 2020) introduces multimodal adaptation gates to inject visual and acoustic information into pre-trained language representations. MulT (Tsai et al., 2019) employs directional cross-modal attention to model long-range interactions across modalities. MMIM (Han et al., 2021) improves multimodal learning through mutual information maximization between modality representations. More recent approaches further explore token-level and structure-aware fusion strategies. TCL-MAP (Zhou et al., 2024) performs token-wise cross-modal alignment to improve multimodal semantic interaction, while SDIF (Huang et al., 2024) enhances multimodal fusion through disentangled semantic representations.

Despite their effectiveness, existing multimodal intent understanding methods mainly focus on improving cross-modal interaction and feature aggregation (Tsai et al., 2019; Rahman et al., 2020; Zhang et al., 2023, 2025a). Most approaches implicitly assume modality consistency and uniformly fuse multimodal information (Tsai et al., 2019; Rahman et al., 2020). However, in real-world open-world environments, visual and acoustic signals are often noisy, ambiguous, or semantically inconsistent with textual content.

In contrast, our work explicitly models modality conflict as a reliability signal and dynamically suppresses unreliable modality evidence during multimodal representation learning.

2.2 Open-World and OOD Intent Detection

Compared with conventional closed-set intent classification, open-world settings introduce substantial uncertainty due to unseen semantic distributions and distribution shifts (Zhang et al., 2023, 2025a; Tang et al., 2024). This challenge becomes more pronounced in multimodal environments, where unreliable visual or acoustic signals may further increase representation ambiguity.

Existing OOD detection methods mainly estimate prediction uncertainty through confidence-based or distance-based scoring strategies. Early approaches commonly employ Maximum Softmax Probability (MSP) (Hendrycks and Gimpel, 2017) as a simple confidence estimator. Subsequent studies explore more robust scoring mechanisms, including energy-based scoring, K-nearest neighbor (KNN) estimation (Zhou et al., 2022), and Mahalanobis Distance (MD) (Podolskiy et al., 2021), which leverage feature-space geometry for OOD discrimination.

Recent studies have extended open-world learning to multimodal intent understanding. MIntOOD (Zhang et al., 2024a) introduces pseudo-OOD sampling and advanced scoring strategies to improve multimodal OOD detection performance. Although these approaches improve uncertainty estimation at the scoring level, they mainly rely on classifier outputs or feature-space statistics while paying limited attention to the reliability of multimodal representations themselves.

In contrast, our work focuses on trustworthy multimodal representation learning under modality conflict. Rather than only improving post-hoc OOD scoring, we explicitly model cross-modal disagreement during fusion and learn more reliable multimodal representations for open-world intent understanding.

2.3 Trustworthy Multimodal Representation Learning

Recent advances in reliability-aware AI highlight the importance of robustness, uncertainty estimation(Bezirganyan et al., 2025), and reliable representations under distribution shifts (Zhang et al., 2025b). In multimodal settings, robust representations are critical because heterogeneous modalities often show varying reliability under noise or open-world conditions (Zhang et al., 2024b).

Existing methods mainly enhance robustness via cross-modal alignment, contrastive learning, or uncertainty-aware optimization (Zhang et al., 2025a; Bezirganyan et al., 2025; Pan and Meng, 2024; Long et al., 2026). Common strategies include estimating modality confidence scores or adaptive fusion weights, as well as noise-aware learning and modality dropout to address imbalance. These approaches have shown effectiveness under incomplete or corrupted inputs.

Recent work further emphasizes uncertainty estimation in multimodal reasoning (Pan and Meng, 2024; Long et al., 2026), such as modality uncertainty calibration and confidence-aware fusion to downweight noisy modalities. However, most existing methods treat unreliable signals primarily as optimization noise or calibration targets, without explicitly modeling how cross-modal disagreement shapes representation geometry under open-world uncertainty.

Unlike prior work, our approach explicitly treats modality conflict as a representation reliability signal. We quantify cross-modal disagreement through attention uncertainty and dynamically suppress unreliable modality evidence during fusion. By focusing on representation geometry, our method achieves clearer ID/OOD separation in open-world settings, going beyond merely improving classifier confidence.

3 Methodology

In open-world intent understanding, models must classify in-domain (ID) intents while rejecting out-of-distribution (OOD) samples. To address representation entanglement caused by cross-modal disagreement, we propose Conflict-Aware Representation Fusion (CARF). As conceptually illustrated in Figure 2, CARF reshapes the feature space through four sequential modules.

3.1 Conflict Estimation

Heterogeneous inputs are first encoded into d-dimensional sequences $\mathbf{V}, \mathbf{A}$ and global representations $\mathbf{t}, \mathbf{v}, \mathbf{a}$ using pretrained encoders. Rather than blindly aggregating these features, CARF actively estimates epistemic uncertainty. Using the textual representation $\mathbf{t}$ as the query, we compute cross-modal attention probabilities over the non-textual sequence (e.g., visual features $\mathbf{V}$): $\mathbf{P}_v = \text{Softmax}(\mathbf{t}\mathbf{V}^\top/\sqrt{d})$. Modality conflict is then explicitly quantified using the Shannon entropy of this attention distribution:

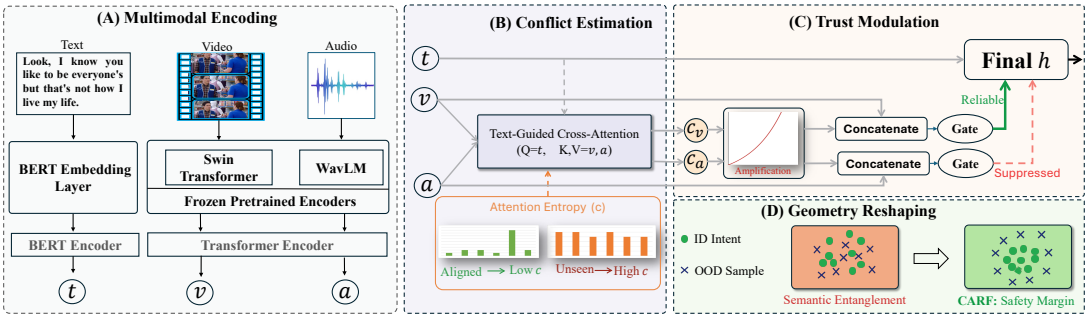

Figure 2: **The overall conceptual framework of Conflict-Aware Representation Fusion (CARF). (A)** Heterogeneous multimodal signals are mapped into a shared feature space. **(B)** Rather than blindly fusing features, CARF actively estimates epistemic uncertainty via cross-modal attention entropy. Concentrated attention indicates semantic alignment (low conflict), while scattered attention signifies modality disagreement (high conflict). **(C)** Guided by the exponentially amplified conflict score, the trust modulation mechanism acts as a circuit breaker—preserving reliable modalities while actively suppressing noisy or contradictory signals. **(D)** Consequently, this conflict-aware modulation fundamentally reshapes the representation geometry, transforming a severely entangled feature space into a highly separable manifold with a clear safety margin for robust out-of-distribution (OOD) detection.

$$c_v = -\frac{1}{\log L_v} \sum_{i=1}^{L_v} P_v^{(i)} \log P_v^{(i)}, \qquad (1)$$

where higher entropy ($c_v \to 1$) indicates scattered attention, serving as a natural proxy for semantic disagreement. The acoustic conflict c_a is computed symmetrically. Detailed formulations are provided in Appendix A.

3.2 Trust Modulation & Regularization

To decisively suppress unreliable modalities, we apply non-linear exponential amplification to construct the trust-aware features: $\mathbf{c}_v = [\bar{c}_v, \bar{c}_v^2, e^{\bar{c}_v}]$, where $\bar{c}_v = 1 - c_v$ inversely reflects modality reliability. These features guide lightweight gating networks to generate trust weights $\mathbf{g}_v$ and $\mathbf{g}_a$. The final robust representation $\mathbf{h}$ is fused by suppressing unreliable evidence:

$$\mathbf{h} = \mathbf{t} + \mathbf{g}_v \odot \mathbf{v}_{aligned} + \mathbf{g}_a \odot \mathbf{a}_{aligned}, \qquad (2)$$

where $\mathbf{v}_{aligned}$ and $\mathbf{a}_{aligned}$ are the residual-aligned global features incorporating text-attended semantics (see Appendix A). Furthermore, we introduce entropy-guided noise regularization during training, dynamically injecting Gaussian noise scaled by the conflict score ($\sigma_v \propto c_v$). This acts as a self-paced regularizer, preventing the model from overfitting to spurious alignments.

3.3 Geometry Reshaping

The purified representation $\mathbf{h}$ is optimized via a cosine classifier. During inference, this trust-modulated fusion fundamentally reshapes the representation geometry. This structural purification enables distance-based metrics, specifically Mahalanobis Distance (MD), to achieve highly robust OOD detection without requiring external pseudo-data.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on three benchmark datasets representing diverse multimodal scenarios. MIntRec2.0 (Zhang et al., 2024a) is a large-scale open-world dataset (30 ID classes, 5,700 OOD samples). MELD-DA (Saha et al., 2020) is derived from conversational television interactions, characterized by high semantic incongruity. IEMOCAP-DA (Saha et al., 2020) contains scripted interactions, providing a high-consistency scenario for evaluation. For MELD-DA and IEMOCAP-DA (11 ID classes), the "Others" class is treated as out-of-distribution (OOD). Following standard open-set protocols (Hendrycks and Gimpel, 2017), OOD data appears only during inference. We utilize pre-extracted features (Zhang et al., 2024a, 2025a) from BERT (Chen et al., 2019), Swin Transformer (Liu et al., 2021), and WavLM (Chen et al., 2022). Detailed statistics are in Table 1.

Table 1: Statistics of the benchmark datasets. T, V, and A denote Text, Video, and Audio modalities. The Test set format is presented as (ID Samples / OOD Samples).

Dataset	Modalities	ID Classes	Train	Valid	Test (ID / OOD)
MIntRec2.0	T, V, A	30	6,165	1,106	2,033 / 1,197
MELD-DA	T, V, A	11	6,561	937	1,875 / 615
IEMOCAP-DA	T, V, A	11	6,541	935	1,870 / 70

Evaluation Metrics. Following established settings (Zhang et al., 2024a, 2025a), we evaluate ID classification using Accuracy (Acc) and Macro F1-score (F1). For OOD detection (Lee et al., 2018; Liang et al., 2017; Ryu et al., 2018), we use the Area Under the Receiver Operating Characteristic Curve (AUROC) and the False Positive Rate at 95% True Positive Rate (FPR95) to quantify safety boundaries.

Baselines & Implementation. We benchmark CARF against recent state-of-the-art unimodal, multimodal fusion, and open-world intent detection methods such as SDIF (Huang et al., 2024) and MIntOOD (Zhang et al., 2024a). Detailed descriptions of comparative baselines and hyperparameter implementation settings are provided in Appendix B. All reported experimental results are averaged over five independent runs using varying random seeds (from 0 to 4).

4.2 Main Results

The overall performance comparisons are presented in Table 2. First, when compared to standard multimodal baselines, CARF consistently establishes the lowest FPR95 scores. For instance, while recent approaches like SDIF achieve high ID accuracy (59.30% on MIntRec2.0), they produce an FPR95 exceeding 94%, frequently misclassifying unknown inputs as ID categories.

It is worth noting that standard baselines typically default to MSP. In contrast, CARF fundamentally purifies the representation geometry, unlocking the potential of Mahalanobis Distance (MD). Evaluated via MD, CARF significantly reduces the FPR95 to 57.94% on MIntRec2.0 and 16.28% on IEMOCAP-DA, overcoming the tendency of conventional fusion architectures to over-align conflicting signals. A rigorous apples-to-apples comparison controlling for scoring functions is dedicated to Section 4.3.

Second, we compare CARF against the open-world state-of-the-art method, MIntOOD. Table 2 highlights that MIntOOD relies on the exposure of pseudo-OOD data during the training phase. Operating under a strict zero-OOD exposure setting, CARF demonstrates absolute FPR95 reductions of 25.15% on MIntRec2.0 (83.09% vs. 57.94%) and 24.40% on MELD-DA (72.49% vs. 48.09%) compared to MIntOOD.

Finally, regarding ID intent classification, CARF maintains competitive or superior performance across all benchmarks, achieving state-of-the-art or highly competitive accuracy on both MIntRec2.0 (59.07%) and MELD-DA (64.51%). The results confirm that the integration of the conflict-aware gating mechanism and noise regularization does not degrade the representation of known semantics, ensuring reliable performance in both closed and open-world deployment scenarios.

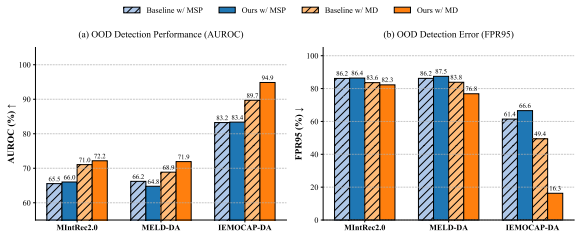

Figure 3: Performance comparison between MSP and Mahalanobis Distance (MD). CARF barely shifts performance when using MSP but drastically unlocks the potential of MD, indicating a geometric reshaping of the multimodal feature space.

4.3 Analysis on Scoring Functions: Why Mahalanobis Distance?

To better understand the geometric properties of the representations learned by CARF, we compare the OOD detection performance using two distinct scoring functions: Maximum Softmax Probability (MSP) (Hendrycks and Gimpel, 2017) and Mahalanobis Distance (MD) (Lee et al., 2018). While MSP is a logit-based confidence estimator, MD operates directly on the representation space, providing a geometric perspective on class separability.

As illustrated in Figure 3, evaluating CARF with the logit-based MSP yields negligible improvements over the baseline. However, switching to the distance-based MD triggers a substantial performance leap, drastically reducing the FPR95 on IEMOCAP-DA from 49.43% to 16.28%. This consistent trend across datasets highlights a critical limitation of logit-based evaluation in multimodal settings: neural networks often map entangled multimodal features to overconfident logits, masking underlying uncertainty.

The significant superiority of MD over MSP substantiates our core design: CARF functions as a structural regularizer rather than a post-hoc logit calibrator. By leveraging entropy-guided noise and conflict gating to penalize modality disagreement, CARF directly purifies the penultimate representation space. It forms compact ID feature clusters while pushing OOD samples to the periphery, en-

Table 2: Main results comparison on MIntRec2.0, MELD-DA, and IEMOCAP-DA. We report ID classification metrics (Acc, F1) and OOD detection metrics (AUROC, FPR95). **Bold** indicates the best results globally, and underlined indicates the second best. The "Extra OOD Data" column specifies whether the method incorporates pseudo-OOD or external samples during training. Scoring functions: $^\diamond$ uses Maximum Softmax Probability (MSP), and ‡ uses Mahalanobis Distance (MD). All reported results are averaged over five independent runs with different random seeds (from 0 to 4).

Method	Extra OOD Data?	MIntRec2.0				MELD-DA				IEMOCAP-DA			
		Acc ↑	F1 ↑	AUROC ↑	**FPR95 ↓**	Acc ↑	F1 ↑	AUROC ↑	**FPR95 ↓**	Acc ↑	F1 ↑	AUROC ↑	**FPR95 ↓**
Standard Multimodal Baselines (Zero OOD Exposure)													
Text-Only$^\diamond$	✗	57.53	51.65	64.61	87.40	63.99	53.63	66.76	84.55	74.83	71.52	81.85	80.45
MAG-BERT$^\diamond$	✗	56.89	43.96	62.76	90.49	64.11	53.78	63.58	90.79	73.81	70.95	77.49	96.15
MulT$^\diamond$	✗	58.57	50.74	63.77	88.20	61.76	54.33	63.32	89.73	74.09	70.47	75.11	91.25
MMIM$^\diamond$	✗	57.94	52.40	64.63	86.92	63.21	**55.23**	59.75	89.92	74.89	71.65	83.40	56.84
TCL-MAP$^\diamond$	✗	58.27	52.81	64.12	87.62	64.27	53.64	65.21	84.33	74.33	71.05	67.71	80.99
SDIF$^\diamond$	✗	**59.30**	**53.04**	61.56	94.12	64.18	53.67	66.13	89.95	**75.02**	71.26	89.38	96.39
Open-World SOTA (Requires Pseudo-OOD Data)													
MIntOOD (MD)‡	✓	58.04	52.49	71.25	83.09	63.76	53.59	76.09	72.49	**75.02**	70.66	**98.14**	**11.72**
Ours (Zero OOD Exposure)													
CARF (Ours)‡	✗	59.07	52.58	**72.17**	**57.94**	**64.51**	53.46	71.93	**48.09**	74.88	70.20	94.86	16.28

Table 3: Ablation study on the structural components of CARF using Mahalanobis Distance for OOD scoring. ID performance is evaluated via Macro F1, and OOD detection via AUROC and FPR95.

Configuration	IEMOCAP-DA			MIntRec2.0		
	F1 ↑	AUROC ↑	FPR95 ↓	F1 ↑	AUROC ↑	FPR95 ↓
Fusion Paradigms						
Add (Baseline)	69.43	89.71	49.43	51.33	71.05	83.59
Concat	71.80	88.79	55.43	52.47	70.23	86.52
MoE	71.88	89.24	29.43	52.38	71.80	82.01
Modality Suppression Mechanisms						
Heuristic (No Gate)	73.07	93.38	30.29	52.00	67.38	84.79
Implicit Gate (No CF)	70.91	94.80	23.43	51.97	71.37	82.94
Regularization						
CARF (w/o Noise)	70.51	93.42	42.57	52.41	71.68	82.91
CARF (Full Model)	70.20	**94.86**	**16.28**	**52.58**	**72.17**	**82.26**

abling distance-based metrics like MD to naturally capture the improved geometric separability without requiring external data.

4.4 Ablation Study

To systematically evaluate the individual contributions of CARF components, we conduct a comprehensive ablation study using Mahalanobis Distance (MD) scoring. Table 3 presents the ID classification (F1) and OOD detection (AUROC, FPR95) results under varying configurations.

Impact of Fusion Paradigms. We first compare CARF against standard content-based routing (Concatenation and Mixture-of-Experts). As shown in Table 3, while increasing parametric complexity via MoE yields marginal ID classification gains (F1: 71.88 on IEMOCAP-DA), it results in suboptimal OOD detection (FPR95: 29.43%). Similarly, Concatenation increases the FPR95 to 55.43% compared to the additive baseline. This indicates that standard content-driven routing overfits to known ID multimodal alignments, degrading open-world separability. Conversely, by modulating features based on epistemic uncertainty (entropy), CARF drastically reduces the FPR95 to 16.28% on IEMOCAP-DA.

Necessity of Explicit Modality Suppression. We isolate specific choices within the gating module. First, replacing the learnable gate with direct scalar multiplication (*Heuristic Weighting*) causes significant instability, dropping the AUROC to 67.38% on MIntRec2.0. Second, utilizing a gating network but omitting the explicitly amplified conflict features (*Implicit Gate*) leaves the FPR95 at 23.43% on IEMOCAP-DA. Integrating the non-linearly expanded conflict features (c, c^2, e^c) into the full CARF model further compresses the FPR95 to 16.28%. This confirms that the exponential term e^c provides a necessary activation gradient, acting as a circuit breaker to decisively shut down unreliable modalities.

Effect of Entropy-Guided Regularization. Finally, removing the dynamic noise injection during training (*CARF w/o Noise*) severely degrades OOD detection, increasing the FPR95 on IEMOCAP-DA from 16.28% to 42.57%. This confirms that injecting Gaussian noise scaled proportionally to modality conflict serves as an essential regularizer. It structurally prevents the model from overfitting to noisy or unaligned features, ensuring the geometric compactness required for distance-based OOD evaluation.

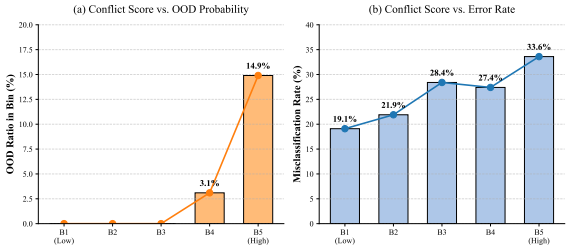

Figure 4: Binning analysis of the conflict score on IEMOCAP-DA. Test samples are partitioned into five equal bins based on conflict severity (Low to High). (a) The ratio of OOD samples surges exclusively in the highest conflict bin. (b) The ID misclassification rate climbs monotonically with conflict severity, validating the score as a reliable trustworthiness indicator.

4.5 Mechanism Analysis: Unveiling the Conflict Score

To understand the physical meaning of the explicitly modeled conflict score c (derived from attention entropy), we investigate its ranking capability and correlation with epistemic uncertainty on the IEMOCAP-DA dataset.

Conflict as an Intrinsic OOD Indicator. We first evaluate whether the raw conflict score naturally captures OOD semantics prior to distance-based evaluation. Using the aggregated raw conflict score as a direct binary classifier yields a high AUROC of **91.84%**. Furthermore, a Top-K capture analysis demonstrates substantial discriminative power: within the top 10% of samples exhibiting the highest conflict, the actual OOD ratio reaches **24.23%**. Compared to the overall random baseline of **3.61%**, this represents a nearly **7-fold enrichment**. These results structurally validate our core hypothesis: unknown intents naturally induce cross-modal disagreement, causing attention dispersion (high entropy) that acts as a natural proxy for epistemic uncertainty.

Binning Analysis for Trustworthiness. To provide a granular view of system reliability, we sort all test samples by their conflict magnitude and partition them into five uniform bins (from Low to High). As illustrated in Figure 4(a), the proportion of OOD samples remains suppressed in lower bins but experiences a sharp, long-tail surge exclusively in the highest conflict bin (B5).

Crucially, extending this binning analysis to the ID classification error rate reveals a stark monotonic correlation (Figure 4(b)). The misclassification rate climbs steadily from 19.1% in the lowest conflict bin (B1) to 33.6% in the highest conflict bin (B5). This confirms that even within the known data distribution, samples with noisy or contradictory multimodal expressions inherently trigger higher conflict scores. By exposing this metric, CARF functions as a highly transparent indicator of reliability, enabling practical NLP systems to selectively flag or reject highly uncertain inputs.

For brevity, we visualize the analysis on IEMOCAP-DA, which exhibits the clearest semantic alignment. Crucially, as detailed in Appendix C, this pronounced long-tail OOD enrichment (i.e., OOD ratios surging exclusively in the highest conflict bin) remains remarkably consistent even on MELD-DA, a dataset heavily characterized by real-world background noise and conversational sarcasm.

4.6 Feature Space Visualization and Geometry Analysis

To visually understand why CARF excels at distance-based OOD detection, we analyze the geometric structure of the learned penultimate representation space on IEMOCAP-DA.

Figure 5 presents a t-SNE projection of the testing features. As shown in Figure 5(a), standard additive fusion produces highly entangled multimodal representations. OOD samples are deeply embedded within the known ID manifolds, explaining why distance metrics fail to distinguish them. In contrast, Figure 5(b) demonstrates the profound geometric reshaping achieved by CARF. Empowered by conflict-aware gating, ID intents form highly compact intra-class clusters, while OOD samples are aggressively pushed toward the peripheral regions of the feature space.

To quantitatively corroborate this phenomenon, Figure 6 visualizes the kernel density estimation (KDE) of the Mahalanobis Distance scores. Under the baseline architecture (Figure 6(a)), the distance distributions exhibit severe overlap, rendering threshold-based OOD discrimination highly unreliable. However, after applying CARF (Figure 6(b)), a drastic topological transformation occurs: ID representations are compressed into denser manifolds (lower distances), while the OOD distribution shifts significantly to the right.

This structural purification creates a distinct *safety margin* between ID and OOD samples. These visual and geometric observations substantiate our core claim: CARF does not merely adjust decision boundaries at the classifier level. Instead, it leverages modality conflict to fundamentally re-

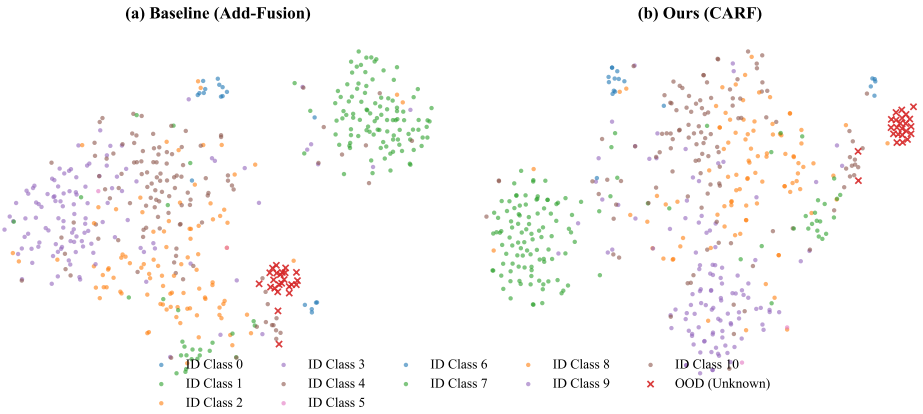

Figure 5: t-SNE visualization of multimodal representations on IEMOCAP-DA. (a) Standard additive fusion suffers from semantic entanglement, where OOD samples (red crosses) sneak into ID clusters. (b) CARF fundamentally purifies the geometry, forming compact ID clusters while robustly marginalizing OOD samples to the absolute periphery.

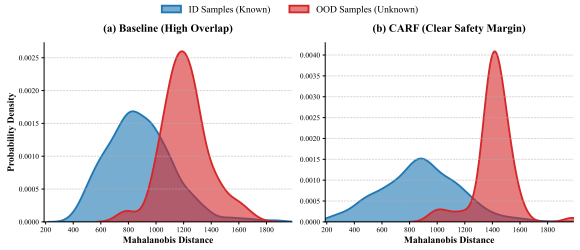

Figure 6: Kernel density estimation of Mahalanobis Distance distributions. (a) Baseline fusion exhibits massive ID/OOD distribution overlap. (b) CARF establishes a distinct safety margin, providing direct numerical evidence for the drastic FPR95 reduction.

structure the representation space, enabling highly reliable open-world deployment.

5 Conclusion

In this paper, we present a paradigm shift in trustworthy multimodal intent understanding by conceptualizing cross-modal conflict as a vital reliability signal rather than mere optimization noise. We propose CARF, a conflict-aware fusion framework that quantifies epistemic uncertainty via cross-attention entropy. In experiments on MIntRec2.0, MELD-DA, and IEMOCAP-DA, operating under a strict zero-OOD exposure setting, CARF maintains competitive in-domain accuracy while achieving massive reductions in False Positive Rates (FPR95) for OOD detection. Our visual and geometric analyses substantiate that CARF does not merely calibrate logit scores, but fundamentally restructures the feature manifold—compressing known intents while marginalizing unseen ones to establish a clear safety boundary.

Limitations

Our work has several limitations. First, our conflict estimation strategy relies asymmetrically on the textual modality as the primary semantic anchor to compute cross-attention entropy. While highly effective for conversational intents, this mechanism presumes that the textual input is relatively intact. If the text itself is severely corrupted (e.g., catastrophic Automatic Speech Recognition (ASR) errors), the reliability of the estimated entropy may degrade. Future work could explore symmetric or modality-agnostic conflict quantification.

Second, our framework utilizes pretrained visual and acoustic features extracted from frozen encoders. We made this design choice to isolate the efficacy of our conflict-aware fusion mechanism and to maintain computational efficiency. However, it restricts the modality-specific encoders from actively adapting to the open-world distribution during training. End-to-end optimization with jointly fine-tuned encoders might yield even tighter feature manifolds.

Third, our experiments mainly focus on static open-world intent detection benchmarks. More challenging real-world settings, such as missing modalities, continual intent evolution, streaming interactions, or severe modality corruption, are not explored in this work. Investigating robust multimodal representation learning under these scenarios remains an important direction for future research.

We provide detailed boundary case analysis in

Appendix D to further illustrate the behavior and limitations of our conflict modeling under realistic multimodal ambiguity.

References

Grigor Bezirganyan, Sana Sellami, Laure Berti-Equille, and Sébastien Fournier. 2025. Multimodal learning with uncertainty quantification based on discounted belief fusion. In *AIStats The 28th International Conference on Artificial Intelligence and Statistics*.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

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.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. 2024. Sdif-da: A shallow-to-deep interaction framework with data augmentation for multimodal intent detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10206–10210. IEEE.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Dong Yu Long, Shuai Wang, Sabariah Md Rashid, and Xiao Tao Lu. 2026. Artificial intelligence in higher education: A systematic review of its impact on student engagement and the mediating role of teaching methods. In *Frontiers in Education*, volume 10, page 1648661. Frontiers.

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

Yawen Ouyang, Jiasheng Ye, Yu Chen, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Energy-based unknown intent detection with data manipulation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2852–2861, Online. Association for Computational Linguistics.

Qiuyu Pan and Zuqiang Meng. 2024. Hybrid uncertainty calibration for multimodal sentiment analysis. *Electronics*, 13(3):662.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13675–13682.

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, Online. Association for Computational Linguistics.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718.

Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multimodal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, Online. Association for Computational Linguistics.

Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. 2024. Contextual augmented global contrast for multimodal intent recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26953–26963.

Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. 2024. Learning geometry-aware representations for new intent discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5641–5654.

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 conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 167–174, Online. Association for Computational Linguistics.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, jinyue Zhao, Wenrui Li, and Yanting Chen. 2024a. 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.

Hanlei Zhang, Hua Xu, Shaojie Zhao, and Qianrui Zhou. 2023. Learning discriminative representations and decision boundaries for open intent detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1611–1623.

Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. 2025a. Multimodal classification and out-of-distribution detection for multimodal intent understanding. *IEEE Transactions on Multimedia*.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. 2024b. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 49279–49383.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. 2025b. Robust multimodal large language models against modality conflict. In *Forty-second International Conference on Machine Learning*.

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.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. Knn-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141.

A Detailed Methodological Formulations

This appendix provides comprehensive mathematical formulations and implementation details for the proposed Conflict-Aware Representation Fusion (CARF) framework, complementing the concise presentation in Section 3.

A.1 Multimodal Feature Encoding

Given a multimodal utterance $X = \{X^t, X^v, X^a\}$, we first extract modality-specific representations using pretrained encoders.

Textual Modality. The textual input is encoded using a pretrained BERT encoder (Chen et al., 2019). We use the hidden representation of the `[CLS]` token as the global semantic representation:

$$\mathbf{t} \in \mathbb{R}^d. \tag{3}$$

Visual and Acoustic Modalities. Frozen pretrained Swin Transformer (Liu et al., 2021) and WavLM (Chen et al., 2022) extract visual/acoustic features offline. Only lightweight Transformers are optimized during CARF training, isolating conflict-aware learning from large-scale pretraining.

$$\mathbf{V} = \text{Transformer}_v(X^v), \quad \mathbf{A} = \text{Transformer}_a(X^a), \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{L_v \times d}$ and $\mathbf{A} \in \mathbb{R}^{L_a \times d}$ represent modality-specific temporal sequences.

The first token of each sequence is used as the global modality representation:

$$\mathbf{v} = \mathbf{V}_0, \quad \mathbf{a} = \mathbf{A}_0. \tag{5}$$

A.2 Text-Guided Cross-Modal Alignment

To estimate modality reliability before fusion, CARF first performs text-guided cross-modal alignment.

For the visual modality, the textual representation serves as the query, while the visual sequence acts as keys and values:

$$\mathbf{P}_v = \text{Softmax}\left(\frac{\mathbf{t}\mathbf{V}^\top}{\sqrt{d}}\right), \tag{6}$$

where $\mathbf{P}_v \in \mathbb{R}^{1 \times L_v}$ denotes the cross-modal attention distribution.

The attended visual representation is computed as:

$$\hat{\mathbf{v}} = \mathbf{P}_v \mathbf{V}. \tag{7}$$

To preserve the original modality semantics while incorporating text-guided alignment information, we adopt a residual integration strategy:

$$\mathbf{v}_{aligned} = \text{LayerNorm}\left(\mathbf{v} + \text{Dropout}(\mathbf{W}_v^{align}\hat{\mathbf{v}})\right). \tag{8}$$

The acoustic modality is processed symmetrically:

$$\mathbf{P}_a = \text{Softmax}\left(\frac{\mathbf{t}\mathbf{A}^\top}{\sqrt{d}}\right), \tag{9}$$

$$\hat{\mathbf{a}} = \mathbf{P}_a \mathbf{A}, \tag{10}$$

$$\mathbf{a}_{aligned} = \text{LayerNorm}\left(\mathbf{a} + \text{Dropout}(\mathbf{W}_a^{align}\hat{\mathbf{a}})\right). \tag{11}$$

The textual representation is normalized directly:

$$\mathbf{t}_{aligned} = \text{LayerNorm}(\mathbf{t}). \tag{12}$$

Diffuse attention distributions indicate that textual semantics fail to consistently align with modality-specific temporal patterns, which we interpret as a proxy for cross-modal uncertainty.

A.3 Conflict-Aware Uncertainty Estimation

CARF estimates modality uncertainty using the entropy of the cross-modal attention distributions.

For the visual modality, the normalized attention entropy is computed as:

$$c_v = -\frac{\sum_{i=1}^{L_v} P_{v,i} \log(P_{v,i} + \epsilon)}{\log(L_v)}, \tag{13}$$

where ϵ is a small numerical constant for stability.

Similarly, the acoustic uncertainty score is:

$$c_a = -\frac{\sum_{i=1}^{L_a} P_{a,i} \log(P_{a,i} + \epsilon)}{\log(L_a)}. \tag{14}$$

Higher entropy indicates that the textual representation attends diffusely across modality-specific temporal regions, reflecting increased cross-modal uncertainty.

Importantly, these uncertainty scores are not treated as direct OOD indicators. Instead, they serve as reliability-aware signals for adaptive feature purification during fusion.

A.4 Trust-Aware Gated Fusion

To dynamically regulate unreliable modality information, CARF employs lightweight trust-aware gating networks.

We first convert uncertainty scores into confidence-like signals:

$$\bar{c}_v = 1 - c_v, \quad \bar{c}_a = 1 - c_a. \tag{15}$$

To enhance sensitivity to reliability variations, we construct amplified confidence descriptors:

$$\mathbf{c}_v = [\bar{c}_v, \bar{c}_v^2, e^{\bar{c}_v}], \tag{16}$$

$$\mathbf{c}_a = [\bar{c}_a, \bar{c}_a^2, e^{\bar{c}_a}]. \tag{17}$$

The visual trust gate is computed as:

$$\mathbf{g}_v = \sigma\left(\mathbf{W}_v^g[\mathbf{v}_{aligned}; \mathbf{c}_v]\right), \tag{18}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function.

The acoustic gate $\mathbf{g}_a$ is obtained analogously.

Higher modality uncertainty leads to smaller confidence descriptors and thus stronger suppression by the trust gating module.

A.5 Conflict-Aware Noise Regularization

To improve representation robustness and prevent overfitting to unreliable modality evidence, CARF injects adaptive Gaussian noise during training.

For the visual modality:

$$\tilde{\mathbf{v}} = \mathbf{v}_{aligned} + \epsilon_v, \tag{19}$$

where

$$\epsilon_v \sim \mathcal{N}(0, \sigma_v^2). \tag{20}$$

The noise scale is dynamically adjusted according to modality uncertainty:

$$\sigma_v = (c_v + \alpha)\beta, \tag{21}$$

where α and β are scaling hyperparameters.

The acoustic modality is regularized similarly.

This uncertainty-conditioned perturbation encourages CARF to learn smoother and more robust multimodal manifolds under unreliable cross-modal conditions.

A.6 Final Representation Fusion

The final multimodal representation is computed as:

$$\mathbf{h} = \mathbf{t}_{aligned} + \mathbf{g}_v \odot \mathbf{v}_{aligned} + \mathbf{g}_a \odot \mathbf{a}_{aligned}, \tag{22}$$

where $\odot$ denotes element-wise multiplication. *Note: During the training phase, the noise-injected features $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{a}}$ are utilized in place of $\mathbf{v}_{aligned}$ and $\mathbf{a}_{aligned}$ to enforce entropy-guided regularization.*

The textual modality serves as the semantic anchor during trust-aware fusion, while visual and acoustic modalities are adaptively modulated according to their estimated reliability.

The fused representation is further refined via a projection layer:

$$\mathbf{h}_{final} = \mathbf{W}_f \mathbf{h}. \tag{23}$$

A.7 Intent Classification and OOD Scoring

After trust-aware fusion, CARF produces the final multimodal representation:

$$\mathbf{h}_{final} = \mathbf{W}_f \mathbf{h}, \tag{24}$$

where $\mathbf{W}_f$ denotes the final projection layer.

Cosine Intent Classifier. To encourage compact hyperspherical feature geometry, we employ a cosine classifier for intent recognition.

The fused representation is first ℓ_2-normalized:

$$\hat{\mathbf{h}} = \frac{\mathbf{h}_{final}}{\|\mathbf{h}_{final}\|_2}. \tag{25}$$

Similarly, each learnable class prototype is normalized:

$$\hat{\mathbf{w}}_k = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2}. \tag{26}$$

The classification logits are computed using scaled cosine similarity:

$$z_k = s \cdot \hat{\mathbf{h}}^\top \hat{\mathbf{w}}_k, \tag{27}$$

where s is a learnable scaling factor.

The final intent probability distribution is obtained via softmax:

$$p(y = k|X) = \frac{\exp(z_k)}{\sum_j \exp(z_j)}. \tag{28}$$

This normalized cosine geometry encourages features from the same intent class to form compact manifolds while maintaining angular separability across different intents.

OOD Detection via Normalized Mahalanobis Distance. For open-world intent understanding, we estimate OOD uncertainty using Mahalanobis Distance (MD) in the normalized feature space.

Given ID training features $\{\hat{\mathbf{h}}_i\}_{i=1}^{N}$, the centroid of each intent class k is computed as:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:y_i=k} \hat{\mathbf{h}}_i. \tag{29}$$

To improve covariance stability, a shared covariance matrix is estimated from all centered ID features:

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_i (\hat{\mathbf{h}}_i - \boldsymbol{\mu}_{y_i})(\hat{\mathbf{h}}_i - \boldsymbol{\mu}_{y_i})^\top. \tag{30}$$

For a test sample, the normalized Mahalanobis distance is computed as:

$$d_{MD}(\hat{\mathbf{h}}) = \min_k (\hat{\mathbf{h}} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\hat{\mathbf{h}} - \boldsymbol{\mu}_k). \tag{31}$$

Following standard OOD evaluation conventions, the negative Mahalanobis distance is used as the final confidence score:

$$s_{OOD} = -d_{MD}(\hat{\mathbf{h}}), \tag{32}$$

where larger values indicate higher in-distribution confidence.

Since both the cosine classifier and MD scoring operate in the same normalized feature space, CARF promotes geometrically consistent decision boundaries for both ID recognition and OOD uncertainty estimation.

B Extended Experimental Setup

This section provides comprehensive details regarding the comparative baselines and implementation hyperparameters supplementing Section 4.1. All datasets and pretrained models used in this work are publicly available for research purposes under their respective licenses and terms of use. We only use publicly released benchmark datasets and follow their original anonymization and research-use protocols.

B.1 Comparative Baselines.

We benchmark CARF against a diverse spectrum of unimodal and multimodal methods:

- **Text-Only:** We use the pretrained BERT (Chen et al., 2019) as a strong unimodal baseline.
- **Standard Multimodal Fusion:** We compare against representative architectures including MulT (Tsai et al., 2019) (Cross-modal Transformers), MAG-BERT (Rahman et al., 2020) (Multimodal adaptation gating), MMIM (Han et al., 2021) (Mutual Information maximization), SDIF (Huang et al., 2024) (Shallow-Deep Interaction Fusion), and TCL-MAP (Zhou et al., 2024). These models implicitly assume modality consistency and operate under a strict zero-OOD exposure setting.
- **Open-World SOTA:** We also benchmark against recent open-world state-of-the-art MIntOOD (Zhang et al., 2024a). MIntOOD utilizes a generative approach involving pseudo-OOD samples during the training phase. Comparing against MIntOOD allows us to critically assess the robustness of CARF without relying on external or synthetic OOD exposure.

B.2 Implementation Details

The dimensions of the pre-extracted features are set to $D_T = 768$, $D_V = 1024$, and $D_A = 768$ for the textual, visual, and acoustic modalities, respectively. Following practical multimodal intent understanding settings, visual and acoustic features are extracted offline using pretrained Swin Transformer and WavLM models and stored as fixed feature sequences during training.

The pretrained textual encoder (BERT-base-uncased) contains approximately 110M trainable parameters. Visual and acoustic features are extracted offline using pretrained Swin Transformer and WavLM models, respectively, and these backbone encoders remain frozen during training. On top of the extracted multimodal features, CARF introduces lightweight task-specific components, including shallow Transformer encoders for sequential visual/acoustic refinement, cross-modal alignment layers, trust-aware gating modules, and cosine classifiers. These newly introduced components contain approximately 2–3M additional trainable parameters.

The trust gating functions are implemented as lightweight single-layer gating networks with sigmoid activation:

$$\mathbf{g}_m = \sigma(\mathbf{W}_m[\mathbf{h}_m; \mathbf{c}_m]), \tag{33}$$

where $\mathbf{h}_m$ denotes the aligned modality representation and $\mathbf{c}_m$ denotes the corresponding conflict-aware confidence features.

B.3 Hyperparameter Search and Computing Budget

All models are optimized using AdamW (Loshchilov and Hutter, 2019). We perform a grid search over the learning rate within $\{9 \times 10^{-6}, 1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$. All experiments are conducted on a single NVIDIA RTX 4090D GPU (24GB VRAM) using PyTorch 2.1. Due to the offline feature extraction and lightweight fusion modules, the training process is highly efficient, requiring less than 1.5 GPU hours to complete all 5 random seeds for a single dataset.

Table 4: Hyperparameter configurations, model parameters, and computing budget across datasets.

Configuration / Dataset	MIntRec2.0	MELD-DA	IEMOCAP-DA
Optimization Settings			
Learning Rate	9×10^{-6}	3×10^{-5}	3×10^{-5}
Batch Size	128	128	128
Weight Decay	0.01	0.01	0.01
Dropout Rate	0.2	0.2	0.2
Early Stopping Patience	8 epochs	5 epochs	3 epochs
Typical Convergence	∼20 epochs	∼20 epochs	∼20 epochs
Model and Computing Budget			
Trainable Parameters	∼113M	∼113M	∼113M
GPU Type	1 × NVIDIA RTX 4090D (24GB)		
GPU Hours (per 5 seeds)	< 1.5 hrs	< 1.5 hrs	< 1.5 hrs

C Extended Mechanism Analysis on MELD-DA

To demonstrate the generalizability of our explicitly modeled conflict score as an indicator of epistemic uncertainty, we extend the mechanism analysis (corresponding to Section 4.5) to the MELD-DA dataset.

Unlike the scripted studio interactions in IEMOCAP-DA, MELD-DA is derived from real-world television dialogues (*Friends*) characterized by frequent sarcasm, rapid context shifts, and heavy background noise. Despite this inherent semantic incongruity, the ranking capability of the conflict score remains highly effective at the extremes.

As detailed in our Top-K analysis, the raw conflict score achieves a moderate AUROC of 63.41%. However, within the Top 10% highest-conflict samples, the actual OOD capture ratio reaches **46.59%**, which is nearly double the random overall baseline of 24.70%.

Furthermore, Figure 7 illustrates the binning analysis for OOD probability on MELD-DA. Consistent with our main findings on IEMOCAP-DA, the OOD ratio remains relatively suppressed in the low-to-medium conflict bins (B1 to B4, averaging ∼19%) but experiences a sharp, long-tail escalation to **43.8%** in the highest conflict bin (B5). This confirms that explicitly amplified conflict gating effectively identifies highly uncertain multimodal inputs, pushing them to the periphery of the feature space even in heavily noisy environments.

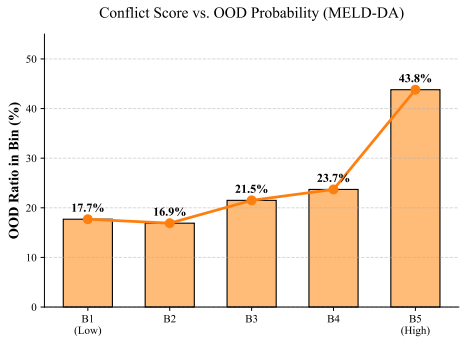

Figure 7: Binning analysis of the conflict score on MELD-DA. The proportion of OOD samples experiences a sharp, long-tail surge exclusively in the highest conflict bin (B5).

D Failure Cases and Boundary Analysis

A core narrative of our work is that cross-modal conflict serves as a pivotal signal to reshape the representation geometry. However, it is crucial to explicitly decouple *modality conflict* from *semantic novelty (OOD)*. While they are highly correlated, as shown in Section 4.5 and Figure 4, they are not strictly equivalent. To provide a transparent boundary analysis of our mechanism, we discuss two edge cases where the raw conflict score diverges from the OOD status, and explain why the combination of CARF and Mahalanobis Distance (MD) robustly handles them.

D.1 Case 1: High-Conflict but In-Distribution (ID)

Scenario: A user expresses a known intent but with modality inconsistency, such as sarcasm (e.g., praising text with a disgusted facial expression) or severe acoustic background noise.
CARF's Behavior: The cross-attention mechanism correctly struggles to align the contradictory signals, resulting in a high attention entropy c. If we were to use the raw conflict score directly as an OOD detector, this ID sample would be falsely rejected. However, CARF utilizes this high conflict to trigger the exponential circuit breaker, aggressively shutting down the unreliable visual/acoustic features via the gating network ($\mathbf{g}_v, \mathbf{g}_a \to 0$).
Outcome: By relying predominantly on the pristine

textual anchor, the final fused representation $\mathbf{h}$ remains tightly bound to its corresponding ID cluster. Consequently, the MD score remains low, correctly classifying it as ID despite the severe modality conflict.

D.2 Case 2: Low-Conflict but Out-of-Distribution (OOD)

Scenario: A user clearly articulates a completely novel intent that was never seen during training (e.g., asking a banking bot to "book a flight"). The textual, visual, and acoustic signals are perfectly aligned and clearly expressed.

CARF's Behavior: Because the modalities are highly consistent, the cross-attention mechanism easily aligns them, yielding a low attention entropy c. The gating network remains open, and the features are fused normally as if it were a clean sample.

Outcome: This highlights that CARF's conflict score is not an omnipotent OOD detector. However, because the semantic content itself is entirely novel, the pretrained encoders map the fused representation to a vacant region in the feature space, far from any established ID centroids. The MD metric effectively captures this geometric distance, successfully rejecting it as an OOD sample.

D.3 Takeaway

These boundary cases illuminate the exact role of CARF. CARF acts as a *feature-space purifier* rather than a *standalone OOD classifier*. By suppressing noisy ID samples (preventing them from bleeding outward) and allowing clean OOD samples to maintain their distinct novel semantics, CARF structurally creates the clear safety margin required for distance-based metrics (like MD) to function flawlessly in the open world.