# A Unified Ambiguity Signal for Calibrated Multimodal Emotion Recognition

**Tahir Ahmad**[1*] , **Raza Mukhtar**[1*] , **Chanjun Park**[2†] , **Yeo Chan Yoon**[1†]
[1]Department of Artificial Intelligence, Jeju National University, Jeju-do, Republic of Korea
[2]School of Software, Soongsil University, Republic of Korea

## Abstract

Multimodal emotion recognition (MER) in conversation is inherently ambiguous, as linguistic, acoustic, and visual cues may support multiple plausible interpretations. Existing approaches often treat ambiguity as prediction uncertainty rather than modeling its structured sources. We propose UAF, an ambiguity-aware framework for MER built around a single unified ambiguity signal. UAF estimates a normalized per-sample Model-based Ambiguity Degree (MAD) from the predicted emotion distribution by measuring its proximity to the maximally ambiguous uniform distribution, and uses it as a shared control signal rather than a post-hoc uncertainty score. This single scalar coordinates three parts of learning: it aligns representation radius with ambiguity, emphasizes ambiguous samples during optimization, and scales OT-based perturbation consistency. The fused evidence is organized into context, speaker, temporal, cross-modal, and residual subspaces that serve as the substrate on which this signal operates. Experiments on IEMOCAP, CMU-MOSEI, and MELD show that UAF yields consistently better calibration and high-ambiguity reliability alongside competitive accuracy and weighted-F1, showing that a single ambiguity signal can be an effective shared training signal for reliable MER.

## 1 Introduction

Multimodal Emotion Recognition (MER) aims to identify human emotions from complementary textual, acoustic, and visual signals. Recent multimodal conversational frameworks have substantially improved cross-modal fusion, contextual reasoning, and conversational representation learning (Hu et al., 2024; Wu et al., 2025a; Shi et al., 2025; Cai et al., 2026). Nevertheless, recognizing emotions in realistic conversational settings remains challenging due to the inherently ambiguous nature of emotional expression. Emotional ambiguity frequently emerges when modality-specific cues conflict, multiple emotions are simultaneously plausible, or contextual and speaker-dependent factors alter emotional interpretation (Wu et al., 2025a). In conversational emotion recognition, these sources are not arbitrary but correspond to dissociable origins of ambiguity documented in affective science. Appraisal theories hold that emotion depends on a context-relative, temporally unfolding evaluation of an utterance rather than on the utterance in isolation, so the same words may support different interpretations across conversational context and dialogue position (Scherer, 2001). This appraisal process is also dynamic: emotions persist across turns and propagate between speakers through inertia and contagion, so the temporal position of an utterance carries affective information (Kuppens et al., 2010; Hatfield et al., 1993). Speaker-specific expression compounds this ambiguity, since emotion research finds no invariant expression prototypes and treats individual expressive variation as the rule rather than the exception (Scherer and Ellgring, 2007; Hamann and Canli, 2004). Ambiguity further arises across modalities. When acoustic, visual, and lexical cues conflict, the perceptual system must resolve the discrepancy, as classically shown by the McGurk effect and its affective analogues in which an incongruent voice shifts the perceived emotion of a face (De Gelder and Vroomen, 2000). A residual source remains beyond these named origins, since pragmatic intent, social role, and cultural background also shape interpretation but lack utterance-level annotations in standard benchmarks. Motivated by these findings, we treat context, speaker, temporal progression, cross-modal interaction, and residual variation as five ambiguity-relevant latent factors that provide a structured inductive bias, without claiming an exhaustive or formally identifiable decomposition. Existing MER

*These authors contributed equally
†Corresponding authors: ycyoon@jejunu.ac.kr, chanjun.park@ssu.ac.kr



*Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2432–2453
November 6-10, 2026 ©2026 Association for Computational Linguistics

approaches have improved multimodal fusion, contextual modeling, and discriminative classification (Shi et al., 2025), while uncertainty-oriented methods often rely on confidence scores, predictive entropy, Bayesian estimation, ensembles, or annotator disagreement. Although these techniques are useful for estimating the reliability of predictions, they generally do not explicitly organize the structured sources from which emotional ambiguity arises. As a result, context dependence, speaker variability, temporal dynamics, and cross-modal conflict may remain entangled in a single representation and be treated as one undifferentiated uncertainty signal. To address these limitations, we propose **UAF**[1] (**U**nified **A**mbiguity **F**ramework), an ambiguity-aware framework for multimodal emotion recognition built around a single unified ambiguity signal that coordinates representation geometry (RAD), optimization, and optimal-transport (OT) robustness. UAF encodes text, audio, and video through modality-specific encoders and cross-modal gated fusion, and organizes the fused representation into five auxiliary subspaces (context, speaker, temporal, cross-modal interaction, and residual variation) that serve as the substrate on which this signal operates. It further computes a normalized per-sample Model-based Ambiguity Degree (MAD) from the predicted emotion distribution by measuring proximity to the maximally ambiguous uniform distribution. This scalar serves as a bounded ambiguity proxy and coordinates three complementary learning mechanisms: radial regularization to align representation radius with ambiguity, ambiguity-weighted optimization to emphasize ambiguous samples, and optimal-transport-based perturbation consistency to improve local robustness. Thus, UAF models ambiguity across latent factors, representation geometry, optimization, and robustness rather than treating it only as post-hoc prediction uncertainty. Our contributions are summarized as follows:

1. We show that a single bounded scalar can simultaneously coordinate three otherwise-separate learning mechanisms: radial regularization, ambiguity-weighted optimization, and ambiguity-scaled optimal-transport consistency, rather than serving only as a post-hoc uncertainty score.

2. We instantiate this control signal as the Model-based Ambiguity Degree (MAD), a bounded scalar computed in one pass and designed to be shared across the three mechanisms rather than to estimate uncertainty.

3. We organize the fused multimodal representation into five auxiliary subspaces that provide a structured inductive bias and serve as the substrate on which MAD operates.

4. We evaluate UAF on IEMOCAP, CMU-MOSEI, and MELD, showing consistent improvements in calibration and high-ambiguity-tier reliability alongside competitive recognition accuracy.

[1] https://github.com/521Tahir/Unified-Ambiguity-Framework-UAF-

## 2 Related Work

### 2.1 Psychological and Linguistic Grounding

Emotional ambiguity in conversation is not random noise; affective science indicates that it arises from identifiable, dissociable sources, which motivates the five latent factors of UAF as a non-exhaustive inductive bias rather than a formally identifiable decomposition. Appraisal theories hold that affect depends on a context-relative, temporally unfolding evaluation of an utterance, so the same words may be interpreted differently across preceding discourse and dialogue position (Scherer, 2001; Moors et al., 2013); in conversation this temporal character appears as emotional inertia and contagion across turns (Kuppens et al., 2010; Hatfield et al., 1993). These observations ground the context factor $z_c$ and temporal factor $z_\tau$. Emotion research further finds no invariant expression prototypes, treating speaker-specific expressive variation as the rule rather than the exception (Scherer and Ellgring, 2007; Hamann and Canli, 2004), which grounds the speaker factor $z_u$. Cross-modal conflict, such as a positive vocal tone with negative lexical content, forces perceptual resolution as in the McGurk effect and its affective analogues (De Gelder and Vroomen, 2000), grounding the modality-interaction factor $z_m$. Remaining influences such as pragmatic intent and cultural background lack utterance-level annotations in standard benchmarks and are absorbed by the residual factor $z_e$.

### 2.2 Ambiguity in Multimodal Emotion Recognition

Multimodal emotion recognition integrates textual, acoustic, and visual signals, and transformer-based

encoders have achieved strong benchmark performance (Lian et al., 2025; Lu and Feng, 2025; Wu et al., 2025b). However, most methods remain discriminative, assigning a single label per utterance and treating ambiguity as noise (Tran et al., 2022; Lian et al., 2025), which limits robustness under mixed emotions or cultural variation (Fang et al., 2022). A parallel line of work models prediction uncertainty using Bayesian networks, ensembles, calibration, or annotator-disagreement soft labels (Prabhu et al., 2023; Matin and Valdenegro-Toro, 2020; Chen et al., 2022), but the relationship between predictive uncertainty and emotional ambiguity is subtle (Hüllermeier and Waegeman, 2021): predictive-distribution dispersion can reflect either limited data and model knowledge or the presence of multiple plausible interpretations shaped by context, intention, and cross-modal conflict (Hong et al., 2025). Rather than claiming to isolate one from the other, UAF uses a dispersion-based scalar as a bounded *proxy* for ambiguity and focuses on how that single signal is propagated through learning. Structured and robustness-oriented representation learning offers tools to separate such heterogeneous factors but remains underexplored in affective computing (Liu et al., 2025; Gamella et al., 2025; Hu et al., 2024), where correlational learning can still entangle modality-specific artifacts with genuine emotional signal (Zhang et al., 2025; Liang et al., 2025). UAF differs by modeling ambiguity through a unified MAD scalar that coordinates representation geometry, optimization, and robustness, supported by a structured latent organization of the fused representation.

### 2.3 Geometric and Optimal-Transport Regularization

Radial and geometric regularization imposes interpretable structure on representations when examples vary in specificity, confidence, or semantic overlap (Araño et al., 2021; Chen et al., 2023; Zheng et al., 2025), which suits emotion labels given their hierarchical and overlapping structure (Zhu et al., 2020). UAF adopts a lightweight Euclidean radial regularizer that links representation radius with prediction-distribution ambiguity, avoiding stronger assumptions about hyperbolic geometry or formal label embeddings. Optimal transport (OT) provides a geometry-aware comparison between empirical distributions and has been used for domain adaptation and robustness analysis across modalities, including point clouds (Zhang et al., 2024) ; (Yang et al., 2025), though its use in ambiguity-aware emotion recognition remains limited. UAF uses OT conservatively as a latent perturbation-consistency regularizer rather than as a causal or interventional mechanism.

## 3 Methodology

We propose **UAF**, an ambiguity-aware multimodal emotion recognition framework organized around a single unified ambiguity signal. UAF learns a fused representation from text, audio, and video, computes a normalized **Model-based Ambiguity Degree** (MAD) from the predicted emotion distribution, and uses it as a unified control signal for radial regularization, ambiguity-weighted optimization, and optimal-transport-based perturbation consistency. The fused representation is organized into five auxiliary subspaces that serve as the substrate on which MAD operates. Figure 1 illustrates the overall architecture.

### 3.1 Problem Formulation

We consider utterance-level multimodal emotion recognition in conversation. For each utterance $i$, the input is $X_i = \{X_i^T, X_i^A, X_i^V\}$, where $X_i^T$, $X_i^A$, and $X_i^V$ denote textual, acoustic, and visual inputs, respectively. The corresponding label is $Y_i \in \{1, \ldots, C\}$, where $C$ is the number of emotion classes. All main experiments follow a single-label multi-class classification setting. The model learns a fused multimodal representation $Z_i = \Phi(X_i) \in \mathbb{R}^{256}$, where $\Phi(\cdot)$ denotes the full multimodal encoder and fusion function.

### 3.2 Multimodal Encoding and Cross-Modal Fusion

For each modality $M \in \{T, A, V\}$, UAF extracts a modality-specific representation using a pretrained encoder and a projection layer, $e_i^M = \mathrm{Proj}_M(f_M(X_i^M; \theta_M)) \in \mathbb{R}^d$, where $f_M(\cdot; \theta_M)$ is the modality encoder and $\mathrm{Proj}_M$ maps its output into a shared embedding space. In our implementation, text, audio, and video are encoded using RoBERTa (Liu et al., 2019), wav2vec2 (Baevski et al., 2020), and ViT-based features (Dosovitskiy et al., 2021), respectively. In our implementation, RoBERTa, wav2vec2, and ViT first encode the sequence-level information within each modality, after which each modality is represented by a projected utterance-level summary vector. The subsequent fusion therefore operates on one summary token per modality rather than

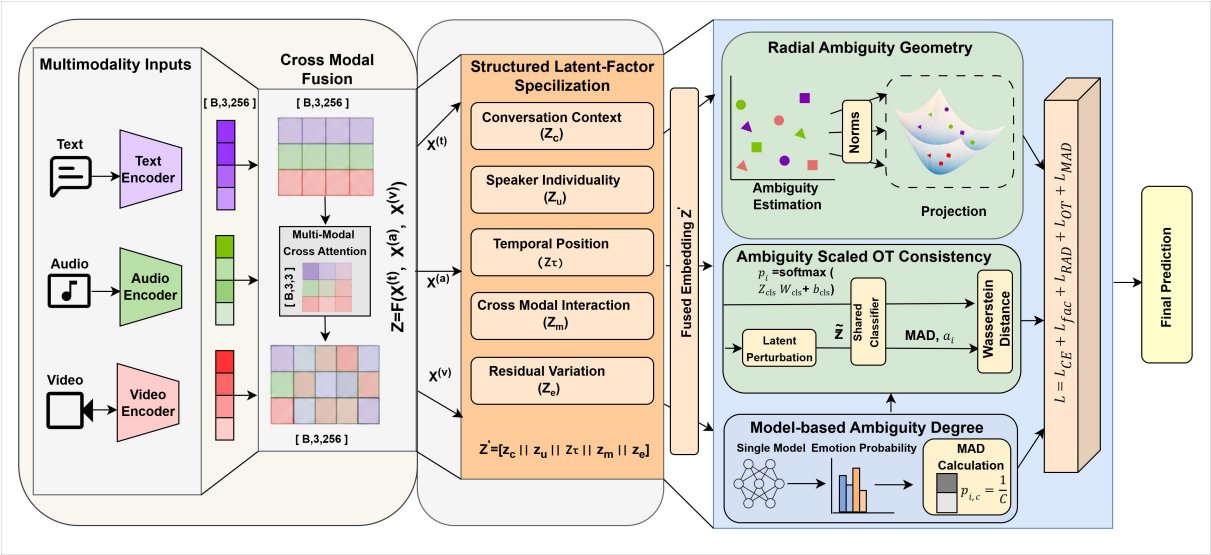



Figure 1: Overall architecture of UAF. Text, audio, and video features are fused and organized into context, speaker, temporal, cross-modal interaction, and residual subspaces that serve as the representational substrate. A single Model-based Ambiguity Degree (MAD), computed from the predicted emotion distribution, coordinates three components of learning: radial representation regularization, ambiguity-weighted optimization, and OT-based perturbation consistency.

on the full token/frame sequences. We apply bidirectional pairwise interactions over the text–audio, text–video, and audio–video summaries, followed by residual connections and layer normalization. Each interaction is followed by a residual connection and layer normalization, yielding $e_i^{T'}$, $e_i^{A'}$, and $e_i^{V'}$. A sample-specific gate $g_i = \sigma(W_g[e_i^{T'}; e_i^{A'}; e_i^{V'}] + b_g) \in \mathbb{R}^3$ then weights each modality, and the gated features are projected into the final 256-dimensional fused representation $Z_i = \mathrm{Proj}_{fuse}([g_{i,T}e_i^{T'}; g_{i,A}e_i^{A'}; g_{i,V}e_i^{V'}]) \in \mathbb{R}^{256}$ used by the rest of UAF. This summary-level design reduces fusion complexity but does not explicitly model fine-grained token-, frame-, or time-level cross-modal alignment. Importantly, UAF's ambiguity-aware components operate on the resulting fused representation $Z$ rather than on encoder internals; thus, sequence-level fusion backbones can replace the current fusion module without changing MAD, RAD, factor organization, or OT consistency. The full fusion architecture is detailed in Appendix A.1. These encoders are representative modality-specific backbones chosen for comparability with prior work. UAF is architecture-agnostic, since all ambiguity-aware components operate on the fused representation $Z$ and its factorized subspaces rather than on encoder internals. Any encoder that produces a fixed-dimensional embedding in the shared space can therefore be substituted without altering the MAD computation, radial regularization, ambiguity-weighted optimization, or OT-based perturbation consistency. We leave a study of stronger backbones to future work.

### 3.3 Structured Latent-Factor Organization

Following the psychological and linguistic grounding in Section 2.1, UAF organizes the fused representation into five auxiliary subspaces corresponding to context, speaker, temporal, modality-interaction, and residual variation. Detailed factor definitions are reported in Appendix A. These subspaces provide a structured inductive bias for ambiguity-aware learning.

### 3.4 Factor Supervision and Decorrelation

UAF guides each subspace with a weak auxiliary objective that encourages it to encode the corresponding source of variation. The context factor is aligned with the mean-pooled fused representation of preceding utterances, $L_{ctx} = \frac{1}{B}\sum_i(1 - \cos(z_{i,c}, \bar{z}_{ctx,i}))$, where $B$ is the mini-batch size. The speaker factor is supervised by speaker-identity classification, $L_{spk} = \frac{1}{B}\sum_i \mathrm{CE}(W_u z_{i,u} + b_u, s_i)$, using the available dialogue or source identifier as weak supervision when speaker labels are absent. The temporal factor regresses the normalized turn position, $L_\tau = \frac{1}{B}\sum_i(w_\tau^\top z_{i,\tau} + b_\tau - \hat{\tau}_i)^2$, where $\hat{\tau}_i$ is the normalized utterance index within its dialogue. The modality-interaction factor regresses the average pairwise embedding-space cosine sim-

ilarity $q_{modal,i} = \frac{1}{3}[\cos(e_i^T, e_i^A) + \cos(e_i^T, e_i^V) + \cos(e_i^A, e_i^V)]$ among the projected modalities representations, $L_{modal} = \frac{1}{B}\sum_i (w_m^\top z_{i,m} + b_m - q_{modal,i})^2$. We use q modal,i as a lightweight, computable proxy for cross-modal interaction rather than as a direct measure of emotional agreement or conflict between modalities; representational similarity in embedding space does not necessarily correspond to consistency of emotional content, since encoders for different modalities capture different aspects of the signal. To discourage redundant linear encoding across factors, we add a lightweight decorrelation penalty

$$L_{indep} = \frac{1}{B|\mathcal{P}|}\sum_{i=1}^{B}\sum_{(r,s)\in\mathcal{P}} \left(\tilde{z}_{i,r}^\top \tilde{z}_{i,s}\right)^2, \quad (1)$$

where $\mathcal{P}$ is the set of unordered factor pairs and $\tilde{z}_{i,r}, \tilde{z}_{i,s}$ are normalized factor vectors truncated to their shared minimum dimension. The full factor-supervision loss is $L_{fac} = L_{ctx} + L_{spk} + L_\tau + L_{modal} + \lambda_{indep} L_{indep}$.

### 3.5 Model-Based Ambiguity Degree

The classifier maps $Z_{cls,i}$ to a softmax emotion distribution $p_i = \mathrm{softmax}(W_{cls} Z_{cls,i} + b_{cls}) \in \mathbb{R}^C$, where $p_{i,c}$ is the predicted probability of class $c$ and $\sum_{c=1}^C p_{i,c} = 1$. We define the MAD as

$$a_i = 1 - \frac{1}{2(1-\frac{1}{C})}\sum_{c=1}^{C}\left|p_{i,c} - \frac{1}{C}\right|. \quad (2)$$

This scalar measures proximity to the maximally ambiguous uniform distribution. The normalization term $2(1-\frac{1}{C})$ is the maximum possible $\ell_1$ distance between a categorical distribution and the uniform distribution, so $a_i \in [0,1]$. Larger values indicate less concentrated predictions and therefore higher model-based ambiguity. MAD is used as a bounded ambiguity proxy rather than as a ground-truth measure of human ambiguity. We adopt the $\ell_1$-to-uniform form rather than entropy because MAD serves as a control signal in three roles, which require a quantity bounded in $[0,1]$ and comparable across datasets. Entropy ranges over $[0, \log C]$ and is scale-dependent on $C$, which would make radial targets and perturbation scales inconsistent across the 4-, 6-, and 7-class datasets without per-dataset renormalization; the class-count-normalized form lets the same $\alpha$ and $\beta$ operate across all three.

### 3.6 Radial Regularization and Ambiguity-Weighted Optimization

To encode ambiguity in representation geometry, UAF aligns the radius of each classifier representation with its ambiguity score. To keep the radial target consistent with the batch-normalised radius, we first centre each sample's MAD score relative to the batch mean, $\bar{a} = \frac{1}{B}\sum_i \mathrm{sg}(a_i)$ and $\tilde{a}_i = \mathrm{sg}(a_i) - \bar{a}$. We then define the normalized radius $r_i = \|Z_{cls,i}\|_2 / (\frac{1}{B}\sum_j \|Z_{cls,j}\|_2 + \epsilon)$, where $\epsilon$ is a small numerical constant, and align it with the centred ambiguity score:

$$L_{RAD} = \frac{1}{B}\sum_{i=1}^{B}\left(r_i - (1 + \alpha\,\tilde{a}_i)\right)^2, \quad (3)$$

where $\alpha$ controls the target spread and $\mathrm{sg}(\cdot)$ denotes stop-gradient. Because $\sum_i \tilde{a}_i = 0$ by construction, the net pull on the batch norms is zero, keeping the denominator stable. This encourages high-ambiguity samples ($\tilde{a}_i > 0$) to occupy larger radii than the batch mean while pushing low-ambiguity samples ($\tilde{a}_i < 0$) below it, enforcing a radius ordering that reflects ambiguity. MAD also weights the per-sample classification loss. With $\ell_i = -\log p_{i,Y_i}$ and $L_{CE} = \frac{1}{B}\sum_i \ell_i$, the ambiguity-weighted auxiliary loss is $L_{MAD} = \frac{1}{B}\sum_i \mathrm{sg}(a_i)\,\ell_i$. Thus ambiguous samples receive stronger optimization emphasis while MAD remains a detached control signal.

### 3.7 Ambiguity-Scaled OT Perturbation Consistency

The third use of MAD is to control latent perturbation strength for optimal-transport-based consistency. UAF perturbs the fused representation before factor projection, $\tilde{Z}_i = Z_i + \eta_i$ with $\eta_i \sim \mathcal{N}(0, \sigma_i^2 I_{256})$ and $\sigma_i = \sigma_0(1 + \beta\,\mathrm{sg}(a_i))$, where $\sigma_0$ is the base scale and $\beta$ controls ambiguity-dependent scaling. Over a mini-batch we form empirical distributions $\mu_Z = \frac{1}{B}\sum_i \delta_{Z_i}$ and $\mu_{\tilde{Z}} = \frac{1}{B}\sum_j \delta_{\tilde{Z}_j}$ over the original and perturbed representations, with pairwise transport cost $C_{ij} = \|Z_i - \tilde{Z}_j\|_2^2$. The entropy-regularized transport plan:

$$P^\star = \arg\min_{P\in\Pi(\mu_Z,\mu_{\tilde{Z}})}\left[\langle P, C\rangle + \epsilon_{OT}\sum_{i=1}^{B}\sum_{j=1}^{B} P_{ij}(\log P_{ij} - 1)\right] \quad (4)$$

where $\Pi(\mu_Z, \mu_{\tilde{Z}})$ is the set of valid transport plans with empirical uniform marginals, $\langle P, C\rangle =$

$\sum_{i,j} P_{ij} C_{ij}$, and $\epsilon_{OT}$ is the Sinkhorn regularization coefficient. The OT consistency loss is $L_{OT} = \langle P^{\star}, C \rangle$. OT is used as a geometry-aware perturbation-consistency regularizer. It does not define ambiguity; instead, MAD controls the perturbation scale, and OT encourages stability between original and perturbed latent distributions.

### 3.8 Stability Motivation and Training Objective

The OT loss is motivated by prediction-distribution stability under latent perturbations. If the classifier map $h : \mathbb{R}^{256} \to \Delta^{C-1}$ is $L$-Lipschitz, then

$$W_2(h_{\#}\mu_Z, h_{\#}\mu_{\tilde{Z}}) \leq L\, W_2(\mu_Z, \mu_{\tilde{Z}}), \quad (5)$$

so reducing the latent OT distance bounds the resulting prediction-distribution shift. We use this as motivation, not as a formal generalization guarantee. The final training objective is

$$L = L_{CE} + \lambda_{fac} L_{fac} + \lambda_{RAD} L_{RAD} + \lambda_{OT} L_{OT} + \lambda_{MAD} L_{MAD} \quad (6)$$

where the $\lambda$ terms control the contribution of each objective. Overall, MAD coordinates three roles: it defines the radial target in $L_{RAD}$, weights ambiguous samples in $L_{MAD}$, and scales latent perturbations for $L_{OT}$.

## 4 Results and Discussion

We evaluate UAF on three standard multimodal emotion recognition benchmarks: IEMOCAP (Busso et al., 2008), CMU-MOSEI (Zadeh et al., 2018b), and MELD (Poria et al., 2019). All datasets are treated as multi-class emotion classification tasks and evaluated using accuracy and W-F1, where W-F1 accounts for class imbalance by weighting each class-specific F1 score according to its class support. For fair comparison, all baselines in Table 1 are run using the same splits, encoders, task formulation, and label sets. Additional baselines, dataset details, and supplementary results are provided in Appendices B and C.

### 4.1 Overall Performance

Table 1 compares UAF with representative multimodal emotion recognition baselines across IEMOCAP, CMU-MOSEI, and MELD. To ensure robustness, all reported results are averaged over three random seeds and presented as mean $\pm$ standard deviation. UAF obtains the best average accuracy in our three-seed evaluation, reaching 81.5 $\pm$ 0.9 accuracy and 80.2 $\pm$ 0.3 W-F1 on IEMOCAP, 74.1 $\pm$ 0.5 accuracy and 55.4 $\pm$ 0.2 W-F1 on CMU-MOSEI; the larger gap between CMU-MOSEI accuracy and W-F1 is due to the imbalanced six-class dominant-emotion protocol, and 63.1 $\pm$ 0.2 accuracy and 63.0 $\pm$ 0.4 W-F1 on MELD. These results show that the proposed framework remains effective in multi-party conversational settings. The gains across datasets suggest that the unified ambiguity signal and the structured latent organization provide complementary benefits over standard multimodal fusion baselines. In particular, improvements over representative transformer-based and dynamic-fusion baselines suggest that MAD-guided radial regularization, ambiguity-weighted optimization, and OT-based perturbation consistency benefit recognition across diverse conversational emotion benchmarks.

### 4.2 Ablation Study

Table 2 reports component-level and latent-factor ablations across IEMOCAP, CMU-MOSEI, and MELD. All ablation variants remove the corresponding component while keeping the remaining hyperparameters fixed, without retuning. Removing RAD, OT, or MAD reduces performance relative to the full UAF model, suggesting that the ambiguity-aware objectives provide complementary benefits within the full training configuration. The ablations affect both accuracy and W-F1, but the dominant effect differs across datasets. On IEMOCAP and CMU-MOSEI, removing ambiguity-aware modules causes larger accuracy drops than W-F1 drops, indicating their contribution to overall recognition robustness. On MELD, W-F1 is more sensitive, especially when OT is removed, suggesting that perturbation consistency is useful in the noisier multi-party setting. In particular, removing MAD produces the largest CMU-MOSEI accuracy drop, while removing OT has the largest effect on MELD W-F1. The latent-factor ablations show that removing any subspace lowers performance relative to the full model, indicating that the structured organization contributes to recognition. Overall, UAF benefits from the interaction of representation geometry, ambiguity-weighted optimization, OT-based robustness, and the structured latent organization. Appendix E reports modality-wise and class-wise analyses. We tune the ambiguity-related loss weights ($\lambda_{\mathrm{MAD}}$, $\lambda_{\mathrm{OT}}$, $\lambda_{\mathrm{RAD}}$) on CMU-MOSEI under LOEO evaluation; moderate MAD weighting, light OT con-

| Method | IEMOCAP | | CMU-MOSEI | | MELD | |
|---|---|---|---|---|---|---|
| | Acc. | W-F1 | Acc. | W-F1 | Acc. | W-F1 |
| LF-MFN (Zadeh et al., 2018a) | 79.0 ± 0.6 | 79.4 ± 0.6 | 73.1 ± 0.8 | 54.2 ± 0.8 | 60.6 ± 0.2 | 61.1 ± 0.4 |
| MulT (Tsai et al., 2019) | 80.1 ± 0.4 | 79.1 ± 0.2 | 73.9 ± 0.6 | 54.3 ± 0.4 | 61.2 ± 0.3 | 61.3 ± 0.5 |
| MM-DFN (Hu et al., 2022) | 80.1 ± 0.8 | 79.2 ± 0.7 | 73.9 ± 0.1 | 53.0 ± 0.3 | 59.5 ± 0.9 | 60.3 ± 0.7 |
| EmoCLIP (Foteinopoulou and Patras, 2024) | 70.8 ± 0.9 | 69.2 ± 0.5 | 72.1 ± 0.8 | 54.3 ± 0.3 | 59.3 ± 0.2 | 60.4 ± 0.3 |
| OV-MER (Lian et al., 2025) | 74.7 ± 0.8 | 74.7 ± 0.9 | 72.1± 0.3 | 51.3 ± 0.1 | 61.9 ± 0.5 | 61.8± 0.3 |
| Ct-NET (Lian et al., 2021) | 80.0 ± 0.5 | 79.9 ± 0.5 | 73.2 ± 0.4 | 52.5± 0.3 | 58.2 ± 0.9 | 59.4 ± 0.6 |
| BP-MulT (Moreno-Galván et al., 2025) | 80.4 ± 0.6 | 79.3± 0.5 | 71.9 ± 0.6 | 51.3± 0.4 | 54.2 ± 0.2 | 57.5 ± 0.6 |
| MESCA (Zhang et al., 2025) | 80.1 ± 0.3 | 79.4 ± 0.2 | 71.3 ± 0.2 | 46.3 ± 0.1 | 60.1 ± 0.7 | 59.1 ± 0.9 |
| **UAF (Ours)** | **81.5 ± 0.9** | **80.2 ± 0.3** | **74.1± 0.5** | **55.4± 0.2** | **63.1 ± 0.2** | **63.0 ± 0.4** |

Table 1: Comparison with representative multimodal emotion recognition baselines across three benchmarks. Results are reported as mean ± standard deviation over three random seeds. Best results are shown in bold.

| Configuration | IEMOCAP | | CMU-MOSEI | | MELD | |
|---|---|---|---|---|---|---|
| | Acc. | W-F1 | Acc. | W-F1 | Acc. | W-F1 |
| w/o RAD | 78.9 | 78.7 | 70.8 | 54.6 | 62.3 | 62.4 |
| w/o OT | 79.9 | 79.3 | 71.6 | 54.4 | 62.9 | 61.1 |
| w/o MAD | 79.5 | 79.4 | 69.4 | 54.3 | 61.7 | 61.9 |
| Latent Factors Ablations | | | | | | |
| w/o $z_c$ | 78.8 | 79.0 | 71.2 | 54.6 | 62.3 | 62.8 |
| w/o $z_u$ | 78.5 | 79.2 | 68.2 | 54.3 | 62.7 | 61.2 |
| w/o $z_\tau$ | 76.1 | 76.5 | 71.2 | 54.6 | 61.3 | 60.7 |
| w/o $z_m$ | 77.6 | 78.1 | 71.2 | 54.6 | 62.4 | 61.7 |
| w/o $z_e$ | 78.1 | 78.3 | 71.3 | 53.9 | 60.9 | 60.1 |
| **Full UAF** | **81.5** | **80.2** | **74.1** | **55.4** | **63.1** | **63.0** |

Table 2: Ablation study of UAF across IEMOCAP, CMU-MOSEI, and MELD. Ablation results are reported from single-seed diagnostic runs. The upper block removes ambiguity-aware learning components, while the lower block removes individual structured latent factors. All ablations keep the remaining hyperparameters fixed without retuning.

sistency, and conservative radial regularization perform best, with the full sweep reported in Appendix E.4.

### 4.3 Calibration Evaluation

Table 3 reports calibration performance on IEMOCAP, MELD, and CMU-MOSEI using Expected Calibration Error (ECE), Brier score, and Negative Log-Likelihood (NLL), where applicable. Calibration is important in ambiguity-aware emotion recognition because models should produce reliable confidence estimates in addition to accurate predictions. UAF achieves the lowest ECE and Brier scores across all datasets and the lowest NLL on IEMOCAP and MELD. On MELD, the ECE and Brier improvements over EmoCLIP are small but consistent with the lowest reported calibration errors. These improvements are especially important in multimodal conversational settings where emotional cues may be conflicting, incomplete, or overlapping across classes. Detailed calibration protocols and metric definitions are provided in Appendix Section F, while additional baseline, Cross Environment Generalization and modality/class-wise analyses are reported in Appendices D and E.

## 5 Comparative Analysis of Ambiguity Proxies

Table 4 compares MAD with entropy, margin, MC-Var, and BALD over three random seeds, each reported at the best validation checkpoint, consistent with the main experiments shown in Table 1. MAD obtains the best mean weighted-F1 and macro-F1 on all three datasets, but the margins are small: on IEMOCAP the proxies differ by well under one standard deviation (e.g., MAD W-F1 $80.20 \pm 0.3$ vs. MC-Var $79.42 \pm 1.24$), and on CMU-MOSEI and MELD MAD and MC-Var are statistically indistinguishable (e.g., MELD W-F1 $63.01 \pm 0.4$ vs. $62.67 \pm 1.15$). Our claim is therefore not that MAD is a categorically stronger uncertainty estimator; several proxies perform comparably. Rather, unlike unbounded scores such as entropy or sampling-based estimates such as MC-Var and BALD, MAD is a bounded, class-count-normalized scalar that is computed in a single forward pass, making it the only comparably-performing proxy that is also directly usable as an efficient inline control signal for radial regularization, ambiguity-weighted opti-

| Method | IEMOCAP | | | MELD | | | CMU-MOSEI‡ | |
|---|---|---|---|---|---|---|---|---|
| | ECE | Brier | NLL | ECE | Brier | NLL | ECE | Brier |
| MulT | 17.06 | 0.394 | 0.972 | 35.47 | 0.915 | 2.234 | 14.62 | 0.164 |
| MM-DFN | 14.18 | 0.372 | 0.785 | 29.80 | 0.845 | 1.908 | 9.70 | 0.154 |
| CTNet | 16.48 | 0.369 | 0.855 | 35.19 | 0.911 | 2.245 | 15.15 | 0.163 |
| BP-MulT | 20.39 | 0.414 | 1.885 | 31.05 | 0.707 | 2.037 | 8.91 | 0.145 |
| EmoCLIP | 15.75 | 0.464 | 0.897 | 20.11 | 0.648 | 1.619 | 6.91 | 0.139 |
| **UAF** | **7.29** | **0.311** | **0.605** | **20.06** | **0.644** | **1.579** | **4.37** | **0.134** |

Table 3: Calibration comparison across datasets. Lower values indicate better calibration. CMU-MOSEI is evaluated under the six-class dominant-emotion softmax protocol used in Table 1. NLL is omitted for CMU-MOSEI only because comparable calibrated logits are unavailable for all baselines under this protocol. Best results are shown in bold.

| Proxy | IEMOCAP | | CMU-MOSEI† | | MELD | |
|---|---|---|---|---|---|---|
| | Mac-F1 | W-F1 | Mac-F1 | W-F1 | Mac-F1 | W-F1 |
| MAD | **80.45** | **80.20** | **46.26** | **55.44** | **45.71** | **63.01** |
| MC-Var | 79.84 | 79.42 | 46.18 | 55.43 | 45.46 | 62.67 |
| BALD | 80.42 | 79.93 | 45.03 | 54.67 | 45.37 | 61.97 |
| Entropy | 80.11 | 79.70 | 44.80 | 54.55 | 44.87 | 62.47 |
| Margin | 79.57 | 79.23 | 45.31 | 54.75 | 44.59 | 62.44 |

Table 4: Ambiguity-proxy comparison across three datasets, averaged over three seeds. MAD gives the best mean performance. †CMU-MOSEI uses the six-class dominant-emotion protocol.

mization, and OT-based perturbation consistency. This is consistent with MAD's role in the method: it is the single bounded scalar that sets the radial target, scales the OT perturbation, and reweights the loss, so a proxy that is both competitive and single-pass is exactly what UAF requires.

## 6 OT Perturbation Behavior Verification

To validate the stability relation in Eq. 5, we measure prediction-distribution shifts using Total Variation (TV) distance under Gaussian latent perturbations with $\sigma \in \{0.01, 0.05, 0.10, 0.20\}$. The TV distance increases monotonically with $\sigma$ while remaining small across all datasets, supporting the Lipschitz-stability motivation of the OT objective and indicating locally stable latent representations; full per-$\sigma$ values and the estimated Lipschitz constants are reported in Table 5. The monotonic increase in TV distance confirms that larger latent perturbations lead to proportionally larger prediction changes. However, the consistently small TV values indicate that the classifier remains locally stable even under stronger perturbations. This behavior supports the role of OT consistency as a robustness mechanism that preserves prediction reliability around ambiguous samples.

| Dataset | $L_{\text{Lip}}$ | $\sigma = 0.01$ | $\sigma = 0.05$ | $\sigma = 0.10$ | $\sigma = 0.20$ |
|---|---|---|---|---|---|
| IEMOCAP | 0.178 | 0.002 | 0.010 | 0.021 | 0.042 |
| CMU-MOSEI | 0.175 | 0.003 | 0.015 | 0.030 | 0.056 |
| MELD | 0.185 | 0.002 | 0.009 | 0.017 | 0.032 |

Table 5: OT perturbation verification under Gaussian latent perturbations ($\sigma$). Values report Total Variation (TV) distance between original and perturbed prediction distributions. $L_{\text{Lip}}$ is the estimated Lipschitz constant of the classifier head.

## 7 Additional Ambiguity-Stratified Results

To avoid circular evaluation using the model's own MAD score, we stratify samples into low-, medium-, and high-ambiguity tiers using dataset-specific proxies external to MAD. IEMOCAP uses annotator-disagreement entropy, CMU-MOSEI uses entropy over discretized AMT annotations, and MELD uses an embedding-based proxy because comparable annotator-disagreement distributions are unavailable.

Tables 6, 7, and 8 show that UAF consistently outperforms EmoCLIP across the evaluated ambiguity tiers. On IEMOCAP, UAF achieves higher accuracy and Macro-F1 and lower ECE in all three tiers, including the high-ambiguity setting (0.6467 Acc., 0.5193 Macro-F1, and 0.2117 ECE). On CMU-MOSEI, UAF similarly achieves the highest weighted accuracy and Macro-F1 and the lowest ECE across all tiers; although the high-tier weighted-accuracy margin is small, UAF retains better Macro-F1 and calibration. On MELD, UAF also performs best across all three metrics and all ambiguity levels, with the largest accuracy advantage appearing in the medium- and high-ambiguity tiers. Overall, the results indicate that UAF maintains stronger recognition and calibration than the comparison model when evaluated using ambiguity measures independent of MAD, including the most ambiguous subsets.

| Tier | Model | Acc. | Macro-F1 | ECE |
|---|---|---|---|---|
| | UAF | **0.8024** | **0.8080** | **0.0870** |
| Low | EmoCLIP | 0.7605 | 0.7598 | 0.1163 |
| | UAF | **0.8144** | **0.8179** | **0.0349** |
| Medium | EmoCLIP | 0.7126 | 0.7209 | 0.1426 |
| | UAF | **0.6467** | **0.5193** | **0.2117** |
| High | EmoCLIP | 0.5928 | 0.4907 | 0.2581 |

Table 6: Dataset-specific ambiguity-stratified evaluation on IEMOCAP. Ambiguity tiers are defined from annotator disagreement and are external to MAD.

| Tier | Model | wAcc. | Macro-F1 | ECE |
|---|---|---|---|---|
| | UAF | **0.9043** | **0.6239** | **0.1528** |
| Low | EmoCLIP | 0.8974 | 0.5888 | 0.1571 |
| | UAF | **0.8039** | 0.4456 | **0.1717** |
| Medium | EmoCLIP | 0.7994 | 0.2903 | 0.4230 |
| | UAF | **0.7421** | 0.2865 | **0.1792** |
| High | EmoCLIP | 0.7418 | 0.2658 | 0.1906 |

Table 7: Dataset-specific ambiguity-stratified evaluation on CMU-MOSEI.

### 7.1 Semantic Cross-Modal Conflict Stress Test

We evaluate trained UAF checkpoints, without retraining, under noisy or emotion-incompatible audio/video and joint A+V conflict. Sensitivity is dataset-dependent: on IEMOCAP, wrong audio, wrong video, and joint A+V conflict reduce accuracy by $16.9 \pm 2.2$, $10.8 \pm 1.4$, and $30.5 \pm 1.6$ points, respectively, whereas MELD and CMU-MOSEI are less affected. Reversed video causes no change due to the order-insensitive utterance-level visual representation and is not interpreted as temporal robustness. Full results and analyses are provided in Appendix G and H.

## 8 Conclusion

This work introduced UAF, an ambiguity-aware framework for multimodal emotion recognition in conversations built around a single unified ambiguity signal. A normalized Model-based Ambiguity Degree, computed from the prediction distribution, coordinates radial regularization, ambiguity-weighted optimization, and optimal-transport perturbation consistency, supported by a structured latent organization of the fused representation. By integrating ambiguity-guided regularization and perturbation consistency, UAF learns more stable and reliable multimodal representations. Experimental results on IEMOCAP, CMU-MOSEI, and MELD demonstrate improvements in recognition performance and calibration, showing the effectiveness of ambiguity-aware learning for multimodal affective understanding.

| Tier | Model | wAcc. | Macro-F1 | ECE |
|---|---|---|---|---|
| | UAF | **0.7333** | **0.4788** | **0.1054** |
| Low | EmoCLIP | 0.7218 | 0.4209 | 0.2359 |
| | UAF | **0.6644** | **0.4814** | **0.1788** |
| Medium | EmoCLIP | 0.6345 | 0.4667 | 0.1877 |
| | UAF | **0.5149** | **0.4270** | **0.2054** |
| High | EmoCLIP | 0.4931 | 0.4241 | 0.2181 |

Table 8: Embedding-proxy ambiguity-stratified evaluation on MELD.

## Limitations

UAF introduces additional computational overhead due to its ambiguity-aware regularization and OT-based consistency objectives. Its robustness evaluation is limited to Gaussian latent perturbations and Total Variation (TV) distance, which may not fully capture structured, adversarial, or real-world distribution shifts. MAD remains a model-derived proxy and therefore may not perfectly reflect human-perceived emotional ambiguity. Similarly, the weakly supervised latent factors should be interpreted as an inductive bias rather than a fully identifiable or interpretable decomposition of ambiguity. UAF also remains sensitive to severe cross-domain modality shifts, particularly when modality characteristics differ substantially from the training data. Future work will investigate stronger domain adaptation, modality-reliability modeling, more diverse perturbation settings, and broader multi-seed validation.

## Ethics Statement

This work uses public datasets and pretrained models under their research-use terms and introduces no new human-subject data. UAF should not be used in high-stakes settings without validation and human oversight, as predictions may reflect cultural and dataset biases. AI assistants were used only for language polishing and editorial support; the authors remain responsible for all scientific content and results.

## Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2026-RS-2026-25615817, AI Star Fellowship Support Program).

## References


Keith April Araño, Carlotta Orsenigo, Mauricio Soto, and Carlo Vercellis. 2021. Multimodal sentiment and emotion recognition in hyperbolic space. *Expert Systems with Applications*, 184:115507.

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

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Yiqiang Cai, Chengyan Wu, Bolei Ma, Bo Chen, Yun Xue, Julia Hirschberg, and Ziwei Gong. 2026. Sure: Synergistic uncertainty-aware reasoning for multimodal emotion recognition in conversations. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11832–11836. IEEE.

Chih-Yao Chen, Tun Min Hung, Yi-Li Hsu, and Lun-Wei Ku. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10947–10958.

Feiyu Chen, Jie Shao, Anjie Zhu, Deqiang Ouyang, Xueliang Liu, and Heng Tao Shen. 2022. Modeling hierarchical uncertainty for multimodal emotion recognition in conversation. *IEEE Transactions on Cybernetics*, 54(1):187–198.

Beatrice De Gelder and Jean Vroomen. 2000. The perception of emotions by ear and by eye. *Cognition & Emotion*, 14(3):289–311.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Xia Fang, Magdalena Rychlowska, and Jens Lange. 2022. Cross-cultural and inter-group research on emotion perception. *Journal of Cultural Cognitive Science*, 6(1):1–7.

Niki Maria Foteinopoulou and Ioannis Patras. 2024. Emoclip: A vision-language method for zero-shot video facial expression recognition. In *2024 IEEE 18th international conference on automatic face and gesture recognition (FG)*, pages 1–10. IEEE.

Juan L. Gamella, Simon Bing, and Jakob Runge. 2025. Sanity checking causal representation learning on a simple real-world system. In *Forty-second International Conference on Machine Learning*.

Stephan Hamann and Turhan Canli. 2004. Individual differences in emotion processing. *Current opinion in neurobiology*, 14(2):233–238.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 6–15.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2025. Aer-llm: Ambiguity-aware emotion recognition leveraging large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE.

Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Lijie Hu, Hasti Seifi, and Jiayuan Xie. 2024. Unimeec: Towards unified multimodal emotion recognition and emotion cause. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5248–5261.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. 2010. Emotional inertia and psychological maladjustment. *Psychological science*, 21(7):984–991.

Zheng Lian, Bin Liu, and Jianhua Tao. 2021. Ct-net: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:985–1000.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, and 1 others. 2025. Ov-mer: Towards open-vocabulary multimodal emotion recognition.

Yachan Liang, Martijn Goudbeek, Agnieszka Konopka, Jiyoun Choi, and Mirjam Broersma. 2025. Investigating cross-cultural vocal emotion recognition with an affectively and linguistically balanced design. *Language and Speech*, page 00238309251318730.

Jianbang Liu, Zhaosheng Xu, Zhongming Liao, Mei Choo Ang, Xiaowei Liu, Kok Weng Ng, Ah-Lian Kor, and Jun Kit Chaw. 2025. A two-tier hierarchical causal model for multimodal conversational emotion reasoning. *Engineering Applications of Artificial Intelligence*, 162:112709.

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

Yuanyuan Lu and Hao Feng. 2025. Multimodal emotion recognition via adaptive high-order transformer network. *Plos one*, 20(10):e0333674.

Maryam Matin and Matias Valdenegro-Toro. 2020. Hey human, if your facial emotions are uncertain, you should use bayesian neural networks! *arXiv preprint arXiv:2008.07426*.

Agnes Moors, Phoebe C Ellsworth, Klaus R Scherer, and Nico H Frijda. 2013. Appraisal theories of emotion: State of the art and future development. *Emotion review*, 5(2):119–124.

Diego Aarón Moreno-Galván, Roberto López-Santillán, Luis Carlos González-Gurrola, Manuel Montes-Y-Gómez, Fernando Sánchez-Vega, and Adrián Pastor López-Monroy. 2025. Automatic movie genre classification & emotion recognition via a biprojection multimodal transformer. *Information Fusion*, 113:102641.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Navin Raj Prabhu, Nale Lehmann-Willenbrock, and Timo Gerkmann. 2023. End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning. *IEEE Transactions on Affective Computing*, 15(2):579–592.

Klaus R Scherer. 2001. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, 92(120):57.

Klaus R Scherer and Heiner Ellgring. 2007. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7(1):113.

Jiandong Shi, Ming Li, Lu Bai, Feilong Cao, Ke Lu, and Jiye Liang. 2025. Match: Modality-calibrated hypergraph fusion network for conversational emotion recognition. In *IJCAI*, pages 6164–6172.

Hélene Tran, Issam Falih, Xavier Goblet, and Engelbert Mephu Nguifo. 2022. Do multimodal emotion recognition models tackle ambiguity? In *Proceedings of the 2nd Workshop on People in Vision, Language, and the Mind*, pages 6–11.

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

ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025a. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6257–6274, Suzhou, China. Association for Computational Linguistics.

Yuezhou Wu, Siling Zhang, and Pengfei Li. 2025b. Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. *Scientific Reports*, 15(1):8855.

Feng Yang, Yichao Cao, Xiu Su, Dan Niu, and Xuanpeng Li. 2025. Counterpc: Counterfactual feature realignment for unsupervised domain adaptation on point clouds. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24760–24769. IEEE.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Xinxin Zhang, Jun Sun, Simin Hong, and Taihao Li. 2024. Amanda: Adaptively modality-balanced domain adaptation for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14448–14458.

Yuqing Zhang, Dongliang Xie, Dawei Luo, and Baosheng Sun. 2025. Modality emotion semantic correlation analysis for multimodal emotion recognition. *Computers and Electrical Engineering*, 126:110467.

Yao Zheng, Guowei Chen, Wenchao Song, Yanchao Liu, and Pengzhou Zhang. 2025. Multimodal hyperbolic embedding and hyperbolic hypergraph fusion for emotion recognition in conversation. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pages 1–8.

Qingyang Zhu, Guanming Lu, and Jingjie Yan. 2020. Valence-arousal model based emotion recognition using eeg, peripheral physiological signals and facial expression. In *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, pages 81–85.

# A Latent Factors

**Conversational Context ($\mathbf{z}_c$).** Prior dialogue turns establish expectations that shape how the current utterance is perceived emotionally. The preceding dialogue window contains the previous $K_c$ utterances, and $\bar{z}_{\text{ctx}}$ is obtained by mean pooling their fused representations.

$$\mathcal{L}_{\text{ctx}} = 1 - \cos\left(\mathbf{z}_c, \bar{\mathbf{z}}_{\text{ctx}}\right) \tag{7}$$

**Speaker Individuality ($\mathbf{z}_u$).** Speaker-specific expression styles substantially influence emotional interpretation, as individual differences in baseline affect and expressivity introduce systematic variability (Busso et al., 2008). The speaker factor is supervised using speaker identity classification:

$$\mathcal{L}_{\text{spk}} = \text{CE}\left(\mathbf{W}_u \mathbf{z}_u + b_u, s_i\right) \tag{8}$$

where $s_i$ denotes the speaker identity for utterance $i$.

**Temporal Position ($z_\tau$).** Emotion evolves across the trajectory of a conversation, where early turns often differ emotionally from later turns. The temporal factor is supervised through normalized turn-index regression:

$$\mathcal{L}_\tau = \left(\mathbf{w}_\tau^\top \mathbf{z}_\tau + b_\tau - \hat{\tau}_i\right)^2 \tag{9}$$

where $\hat{\tau}_i \in [0, 1]$ denotes the normalized position of utterance $i$ within its dialogue.

**Cross-Modal Interaction ($z_m$).** Conflicts between modalities, such as a positive vocal tone accompanying negative lexical content, are a major source of emotional ambiguity (Han et al., 2021), and motivate including a modality-interaction factor in the latent organization. Directly supervising this factor with ground-truth emotional agreement is not possible, since standard MER benchmarks do not annotate which modality carries the "correct" emotion under conflict. We therefore use the average pairwise cosine similarity among the projected modality representations, $q_{modal,i}$, as a cheap, self-supervised proxy target rather than a validated measure of emotional agreement:

$$q_{\text{modal},i} = \frac{1}{3}\Big[\cos(e_i^T, e_i^A) + \cos(e_i^T, e_i^V) + \cos(e_i^A, e_i^V)\Big]. \tag{10}$$

where $e_i^T$, $e_i^A$, and $e_i^V$ denote the projected text, audio, and video representations for utterance $i$, respectively. The modality-interaction loss is then

$$L_{modal} = \frac{1}{B}\sum_{i=1}^{B}\left(w_m^\top z_{i,m} + b_m - q_{modal,i}\right)^2, \tag{11}$$

where $w_m$ and $b_m$ are learnable parameters of the modality-agreement regression head.

We emphasize that $q_{modal,i}$ reflects *embedding-space similarity*, not verified emotional agreement: two modalities can be embedding-dissimilar while emotionally consistent, for instance because text and audio encoders capture different feature types (lexical semantics versus acoustic prosody) that need not align geometrically even when the underlying emotion is the same; conversely, two modalities can be embedding-similar while emotionally conflicting, if superficial representational overlap does not track affective content.

**Residual Variation ($\mathbf{z}_e$).** The remaining latent dimensions absorb variation not captured by the four supervised factors, including pragmatic context, social role, and broader latent influences such as cultural background that lack utterance-level annotations in existing benchmark datasets. The residual factor receives no direct supervision and acts as a bottleneck that discourages factor collapse.

We note that factors such as culture and pragmatics are theoretically important components of $Z$ but are not directly operationalizable in datasets such as IEMOCAP, MELD, or CMU-MOSEI, as these corpora do not provide utterance-level cultural or pragmatic metadata. These sources of variation are therefore partially absorbed by $\mathbf{z}_e$.

## A.1 Multimodal Fusion and Intra-Modal Self-Attention

In the first stage, each projected modality representation is processed independently through an intra-modal self-attention module. The textual, acoustic, and visual feature vectors are denoted as $e_T$, $e_A$, and $e_V$, respectively. Since each modality is represented as a single-token sequence, every modality-specific vector is passed through its own

Transformer encoder layer. This encoder layer follows a pre-normalization structure and uses four attention heads with a feed-forward network dimension of 1024. The transformation can be expressed as follows:

$$e_T \rightarrow \text{TransformerEncoderLayer}(e_T) = e'_T,$$

$$e_A \rightarrow \text{TransformerEncoderLayer}(e_A) = e'_A,$$

$$e_V \rightarrow \text{TransformerEncoderLayer}(e_V) = e'_V,$$

where $e'_T$, $e'_A$, and $e'_V$ are the refined text, audio, and visual representations, each having the dimensionality $[B, 1, 256]$. The purpose of this stage is to enhance each modality within its own feature space before allowing interaction with other modalities. This helps preserve modality-specific information and improves the quality of the representations used in later cross-modal fusion.

### A.2 Bidirectional Cross-Modal Attention

After intra-modal refinement, the model performs bidirectional cross-modal attention among all three modality pairs. This stage allows each modality to exchange information with the remaining modalities. First, the textual and acoustic representations interact with each other, where $e'_T$ attends to $e'_A$ and $e'_A$ attends to $e'_T$. This produces updated representations $e''_T$ and $e''_A$. Similarly, the textual and visual modalities interact, where the updated textual representation attends to the visual representation and the visual representation attends to the textual representation, resulting in $e'''_T$ and $e''_V$. Finally, the acoustic and visual modalities interact, producing the final updated acoustic and visual representations $e'''_A$ and $e'''_V$.

This process can be summarized as:

$$(T, A): \quad e'_T \leftrightarrow e'_A \rightarrow e''_T, e''_A,$$

$$(T, V): \quad e''_T \leftrightarrow e'_V \rightarrow e'''_T, e''_V,$$

$$(A, V): \quad e''_A \leftrightarrow e''_V \rightarrow e'''_A, e'''_V.$$

Each cross-modal update is followed by a residual connection and layer normalization. The residual connection helps retain the original modality-specific information, while layer normalization stabilizes the learned representation. By the end of this stage, each modality representation has incorporated complementary information from the other two modalities. Therefore, the resulting features become more context-aware and better suited for multimodal sentiment or emotion-related reasoning.

### A.3 Mean Pooling and Sentiment-Aware Gating

In the third stage, the final cross-modally updated representations are stacked together to form a multimodal feature tensor. The stacked representation is given by:

$$E = [e'''_T, e'''_A, e'''_V] \in \mathbb{R}^{B \times 3 \times 256}.$$

A sentiment-aware gating mechanism is then applied to dynamically control the contribution of each modality. The three modality representations are concatenated and passed through a learnable gating function followed by a sigmoid activation. The gate is computed as:

$$g = \sigma(W_g[e'''_T; e'''_A; e'''_V] + b_g),$$

where $g \in \mathbb{R}^{B \times 3}$ represents the modality-wise gating scores, $W_g$ and $b_g$ are learnable parameters, and $\sigma$ denotes the sigmoid activation function. The gate values are then expanded to match the feature dimension and multiplied with the stacked modality features:

$$E_g = E \odot g,$$

where $E_g \in \mathbb{R}^{B \times 3 \times 256}$ denotes the gated multimodal representation. This mechanism allows the model to assign higher importance to informative modalities and suppress noisy or less useful modalities for each individual sample. As a result, the model can adaptively focus on the most reliable modality combination depending on the input context.

### A.4 Feature Concatenation and Auxiliary Projection

After applying sentiment-aware gating, the gated modality features are concatenated into a single raw multimodal vector. Since there are three modalities and each modality has a dimensionality of 256, the concatenated representation has a dimension of 768. This operation is represented as:

$$z_{\text{raw}} = \text{Concat}(E_g) \in \mathbb{R}^{B \times 768}.$$

The concatenated vector is then projected into a lower-dimensional fused representation using a linear projection layer:

$$z_{\text{fused}} = \text{Linear}_{768\to256}(z_{\text{raw}}),$$

where $z_{\text{fused}} \in \mathbb{R}^{B\times256}$. This fused vector serves as the unified multimodal representation of the input sample. It combines textual, acoustic, and visual cues into a compact feature space and is subsequently passed to the UAF components for further processing.

### A.5 Factorized Multimodal Representation Learning

The factor projection module is implemented using five independent MLP heads applied in parallel to the same normalized fused representation. Specifically, $z_{\text{fuse}}$ is first passed through a LayerNorm layer. The output is then fed to five separate projection heads $\phi_c, \phi_u, \phi_t, \phi_m, \phi_e$. Each head follows the architecture: The projection head applies a linear transformation, followed by Layer Normalization, ReLU activation, dropout with rate 0.1, and a final linear layer to map the features into the target embedding dimension $d_k$. The five output dimensions are $64, 64, 32, 64, 32$, respectively.

## B Baselines

To evaluate the effectiveness of the proposed framework, we compare UAF against multiple multimodal emotion recognition methods spanning diverse architectures and time periods: LF-MFN (Zadeh et al., 2018a) (memory fusion, 2018), MulT (Tsai et al., 2019) (multimodal transformer, 2019), MM-DFN (Hu et al., 2022) (dynamic fusion, 2022), EmoCLIP (Foteinopoulou and Patras, 2024) (vision-language zero-shot, 2024), OV-MER (Lian et al., 2025) (open vocabulary, 2024), Ct-NET (Lian et al., 2021) (CNN-transformer hybrid, 2021), BP-MulT (Moreno-Galván et al., 2025) (bi-projection transformer, 2025), MESCA (Zhang et al., 2025) (semantic correlation analysis, 2025).

## C Dataset Details

All experiments were conducted using the standard train-test splits defined for each dataset. IEMOCAP contains ∼10k multimodal utterances of 10 speakers in dyadic conversations. CMU-MOSEI is a large-scale, in-the-wild dataset with more than 23k video segments from more than 1k speakers. MELD includes ∼13k utterances from multi-party dialogs in the *Friends* TV series, emphasizing contextual emotion dynamics. CMU-MOSEI supports multiple evaluation formulations. In this work, we adopt the 6-class dominant-emotion classification protocol, where each utterance is assigned a single label corresponding to the emotion with the highest annotation intensity. This setting differs from the commonly reported binary per-dimension evaluation, where each emotion is treated independently. The binary formulation typically yields higher F1 scores (0.65–0.78 range), while the 6-class setting is significantly more challenging due to inter-class competition. All results in Table 1 are computed under the same 6-class protocol for both UAF and all baselines to ensure fair comparison.

## D Cross-Environment Generalization

The LOEO environments are constructed based on speaker-disjoint clusters derived from speaker metadata. We partition the dataset into five speaker-disjoint groups using the provided speaker IDs, ensuring that no speaker appears in both training and test sets within any fold. This is not a random split; it simulates cross-environment generalization where the model must handle unseen speakers with potentially different cultural and individual expression patterns. The Leave-One-Environment-Out (LOEO) protocol uses structured, non-random splits. For IEMOCAP, each session (five total) forms one environment. Since speakers do not overlap across sessions, this yields a speaker-disjoint evaluation. For CMU-MOSEI, environments are defined by video source clusters corresponding to different speakers and channels, capturing varying demographic and cultural contexts. For MELD, environments are constructed from disjoint sets of episodes, preserving full conversational structure and preventing leakage across train and test sets. These splits ensure that evaluation reflects generalization to unseen speakers and contexts rather than random partitioning. Table 9 reports cross-environment performance. The text-only baseline provides a strong reference point; however, incorporating additional modalities yields consistent gains under distribution shift. The full tri-modal configuration achieves the highest weighted F1, although text-only and TV configurations achieve higher accuracy, suggesting a tradeoff between robustness to minority classes and overall accuracy. Under LOEO evaluation, the full tri-modal configuration achieves the highest weighted-F1, while text-

| Modality | Accuracy | Weighted F1 |
|---|---|---|
| Text (T) | 76.86 | 50.52 |
| Text + Audio (TA) | 72.82 | 52.47 |
| Text + Video (TV) | 76.73 | 54.69 |

Table 9: Cross-environment evaluation on CMU-MOSEI using leave-one-environment-out (LOEO) splits. Environments are speaker-disjoint clusters. Results are averaged over five held-out environments.

| Modality | Accuracy | Weighted-F1 |
|---|---|---|
| Text (T) | 49.03 | 37.37 |
| Text + Audio (TA) | 48.73 | 38.84 |
| Text + Video (TV) | **53.84** | 38.72 |
| Text + Audio + Video (TAV) | 43.61 | **39.82** |

Table 10: Cross-dataset cross-environment evaluation from CMU-MOSEI to MELD. Higher values indicate better performance.

only and text-video settings obtain higher accuracy. This suggests that adding all modalities improves class-sensitive robustness under speaker-disjoint evaluation, although accuracy and weighted-F1 respond differently to modality shift.

### D.1 Cross-Dataset Transfer

Table 10 reports cross-dataset generalization results when training on CMU-MOSEI and evaluating on MELD. A notable finding is that the TAV configuration achieves only 43.61% accuracy, lower than TV alone (53.84%). This accuracy degradation reflects negative transfer from the audio modality when the source domain (CMU-MOSEI: diverse YouTube speakers) and target domain (MELD: studio-recorded TV dialogue) exhibit large acoustic gaps. The audio encoder, optimized for YouTube-style speech, produces representations that actively mislead the classifier on MELD's studio-recorded dialogue. Importantly, weighted-F1 still improves (39.82% for TAV vs. 38.72% for TV), indicating that while overall accuracy drops due to majority-class confusion, minority-class recognition benefits from multimodal integration. This result highlights a limitation of the current UAF formulation. Although UAF improves standard benchmark performance, its ambiguity-aware modules do not fully prevent negative transfer when a modality undergoes severe cross-dataset distribution shift. In particular, the audio modality learned from CMU-MOSEI appears poorly aligned with MELD, causing TAV accuracy to degrade relative to TV. Future work should combine UAF with explicit domain adaptation, modality reliability estimation, or environment-aware training objectives.

## E Complete Modality-Subset Analysis

### E.1 Modality Contribution Analysis

Table 11 evaluates all seven subsets of text (T), audio (A), and video (V) under the standard in-domain evaluation protocol. Text provides the strongest unimodal accuracy on all three datasets, whereas video provides the strongest unimodal W-F1 on CMU-MOSEI, indicating that acoustic and visual cues can be particularly useful for class-sensitive recognition. Adding video to text improves W-F1 by 1.94, 3.51, and 4.89 percentage points on IEMOCAP, CMU-MOSEI, and MELD, respectively. Adding video to the text–audio configuration produces further W-F1 gains of 5.98, 10.90, and 2.38 points. These comparisons isolate a contribution from visual evidence that is independent of audio–visual interaction. Similarly, adding audio to text and to text–video improves both accuracy and W-F1 across all three datasets. The full TAV configuration achieves the best performance in every dataset–metric combination. However, the pairwise results are not uniformly monotonic, particularly for CMU-MOSEI W-F1, suggesting that effective recognition depends on complementary three-way fusion rather than on the simple addition of more modalities.

### E.2 Class-wise Performance Analysis

Table 12 reports class-wise accuracy across datasets. On IEMOCAP, accuracy is consistently high across all emotion classes, reflecting the acted and well-segmented nature of the dataset. For CMU-MOSEI, performance varies more substantially, with emotions like fearful and disgust exhibiting lower accuracy due to class overlap in spontaneous speech. On MELD, class-wise accuracy is generally lower and more uneven, particularly for disgust (12.0%), reflecting the conversational and context-dependent nature of the dataset.

### E.3 Implementation details and training configurations

Table 13 summarizes the dataset-specific implementation settings and training configurations used for UAF across IEMOCAP, MELD, and CMU-MOSEI. Although the proposed framework maintains a consistent multimodal architecture with RoBERTa, wav2vec2, and ViT-based encoders, several parameters are adapted according to dataset complexity and task requirements. The choice of backbones is not intrinsic to UAF; these en-

| Modality | IEMOCAP | | CMU-MOSEI | | MELD | |
|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| Text Only | 72.72 | 72.07 | 65.48 | 40.15 | 57.45 | 56.20 |
| Audio Only | 68.72 | 66.36 | 53.63 | 48.15 | 48.10 | 46.05 |
| Video Only | 70.96 | 69.28 | 56.48 | 49.69 | 50.09 | 46.58 |
| Text + Audio | 76.42 | 74.22 | 68.40 | 44.50 | 61.41 | 60.62 |
| Text + Video | 75.06 | 74.01 | 66.95 | 43.66 | 61.64 | 61.09 |
| Audio + Video | 73.75 | 73.41 | 64.77 | 41.66 | 57.74 | 56.38 |
| Text + Audio + Video | **81.5** | **80.2** | **74.1** | **55.4** | **63.1** | **63.0** |

Table 11: Complete modality-subset analysis under the standard in-domain evaluation protocol. All values are reported as percentages. Best results are shown in bold.

| Dataset | Happy | Sad | Angry | Neutral | Fearful | Disgust | Surprised |
|---|---|---|---|---|---|---|---|
| IEMOCAP | 86.0 | 89.0 | 82.0 | 85.0 | – | – | – |
| CMU-MOSEI | 86.5 | 83.2 | 78.9 | – | 71.0 | 72.8 | 76.4 |
| MELD | 51.2 | 30.3 | 58.8 | 77.2 | 29.0 | 12.0 | 55.5 |

Table 12: Class-wise accuracy (%) across IEMOCAP, CMU-MOSEI, and MELD.

coders can be substituted with newer modality encoders as long as their outputs are projected into the shared 256-dimensional fusion space. In particular, MELD employs a larger text encoder and longer training schedule due to its challenging 7-class conversational setting, CMU-MOSEI uses the six-class dominant-emotion formulation with dataset-specific MAD and optimization weights. All datasets share the same fusion strategy, radial regularization settings, and OT-based alignment mechanism, ensuring architectural consistency and fair evaluation. In all main experiments, we set $\beta = 1.0$, the radial target-scale coefficient is set to $\alpha = 1.0$. Training was conducted using distributed learning on NVIDIA RTX A6000 GPUs with dataset-specific adjustments in learning rate, batch size, and regularization to achieve stable convergence and efficient performance.

### E.4 Hyperparameter Sensitivity

Table 14 analyzes the sensitivity of UAF to ambiguity-related loss weights under LOEO cross-environment evaluation on CMU-MOSEI. Moderate MAD weighting provides the strongest weighted-F1, with $\lambda_{\mathrm{MAD}} = 0.3$ achieving the best score in the one-dimensional sweep. OT regularization remains stable across small nonzero values, with $\lambda_{\mathrm{OT}} = 0.05$ yielding the strongest accuracy while maintaining competitive weighted-F1. For radial regularization, small values are preferred: $\lambda_{\mathrm{RAD}} = 0.01$ improves weighted-F1, whereas larger values substantially reduce accuracy. These results indicate that UAF benefits from moderate ambiguity weighting, light OT consistency, and conservative radial regularization.

## F Calibration Protocol and Metric Definitions

For all datasets, we evaluate calibration under the same classification protocol used in Table 1. IEMOCAP and MELD are treated as single-label multi-class emotion classification tasks. CMU-MOSEI is also evaluated as six-class dominant-emotion classification, where each utterance is assigned a single label corresponding to the emotion with the highest annotation intensity. Therefore, CMU-MOSEI uses a six-class softmax probability distribution in our evaluation, not a multilabel sigmoid formulation. We compute Expected Calibration Error (ECE) using 15 equal-width confidence bins. For each prediction, confidence is defined as the maximum softmax probability over the emotion classes. The Brier score is computed using the multi-class probability distribution and the corresponding one-hot ground-truth label. Negative Log-Likelihood (NLL) is reported for IEMOCAP and MELD using the predicted probability assigned to the gold class. For CMU-MOSEI, NLL is omitted only because comparable calibrated logits are not available for all baselines under the same six-class dominant-emotion protocol; this omission is unrelated to the task formulation. UAF achieves the strongest calibration across all three datasets, obtaining the lowest ECE and Brier score among the compared methods on IEMOCAP, MELD, and CMU-MOSEI. On IEMOCAP and MELD, UAF also achieves the lowest NLL. For CMU-MOSEI, NLL is omitted because comparable calibrated logits are unavailable for all baselines under the same six-class dominant-emotion protocol. These results suggest that ambiguity-aware factorization, radial

| Setting | IEMOCAP | MELD | CMU-MOSEI |
|---|---|---|---|
| **Task / Targets** | 4-class emotion classification | 7-class emotion classification | 6-class dominant emotion recognition |
| **Text encoder** | RoBERTa-base (12L, 768H) | RoBERTa-base (12L, 768H) | RoBERTa-base (12L, 768H) |
| **Text parameters** | 124.6M | 124.6M | 124.6M |
| **Audio encoder** | wav2vec2-base | wav2vec2-base | wav2vec2-base |
| **Audio input** | 16kHz waveform, max 4s | 16kHz waveform, max 6s | 16kHz waveform, max 6s |
| **Video encoder** | ViT-base-patch16-224 | ViT-base-patch16-224 | ViT-base-patch16-224 |
| **Video input** | 8 frames, 224×224 | 8 frames, 224×224 | 8 frames, 224×224 |
| **Feature projection** | Linear($d_{enc} \to 256$) + ReLU + Dropout(0.2) | | |
| **Fusion architecture** | Hierarchical multimodal cross-attention with gated fusion + sigmoid fusion gate | | |
| **Fusion output dim** | 256 | 256 | 256 |
| **radial target-scale coefficient** | 1.0 | 1.0 | 1.0 |
| **OT solver** | Log-domain Sinkhorn | Log-domain Sinkhorn | Log-domain Sinkhorn |
| **Sinkhorn $\varepsilon$** | 0.1 | 0.1 | 0.1 |
| **OT perturbation $\sigma$** | 0.05 | 0.05 | 0.05 |
| **OT iterations** | 20 | 20 | 20 |
| **Learning rate** | $1.5 \times 10^{-5}$ | $1.2 \times 10^{-5}$ | $2.0 \times 10^{-5}$ |
| **Optimizer** | AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$) | | |
| **LR schedule** | Warmup 10% + cosine decay | Warmup 12% + cosine decay | Warmup 10% + cosine decay |
| **Weight decay** | 0.01 | 0.02 | 0.01 |
| **Label smoothing** | 0.1 | 0.1 | None |
| **Batch size / GPU** | 6 | 4 | 32 |
| **Max epochs** | 35 | 60 | 60 |
| **Early stopping** | Patience 8 | Patience 15 | Patience 15 |
| **Loss weights** | $\lambda_{fac} = 0.005, \lambda_{RAD} = 0.01, \lambda_{OT} = 0.05, \lambda_{MAD} = 0.3$ | | |
| **Hardware** | 5× NVIDIA RTX A6000 (48GB), DDP | | |
| **Training time** | $\approx$ 2 h | $\approx$ 4 h | $\approx$ 5 h |
| **Inference time** | $<$ 1 min | $\approx$ 2 min | $\approx$ 3 min |
| **Peak GPU memory** | $\approx$ 14 GB | $\approx$ 22 GB | $\approx$ 14 GB |

Table 13: Dataset-specific implementation details and training configurations for UAF.

| Component | $\lambda$ | Acc. (%) | WF1 (%) |
|---|---|---|---|
| MAD ($\lambda_{MAD}$) | 0.0 | 71.31 | 54.39 |
| MAD ($\lambda_{MAD}$) | 0.1 | 72.83 | 53.99 |
| MAD ($\lambda_{MAD}$) | 0.3 | 75.95 | 55.58 |
| MAD ($\lambda_{MAD}$) | 0.5 | 73.19 | 55.30 |
| MAD ($\lambda_{MAD}$) | 0.7 | 72.76 | 55.35 |
| OT ($\lambda_{OT}$) | 0.00 | 73.19 | 55.30 |
| OT ($\lambda_{OT}$) | 0.01 | 74.74 | 55.01 |
| OT ($\lambda_{OT}$) | 0.03 | 74.99 | 54.78 |
| OT ($\lambda_{OT}$) | 0.05 | 75.83 | 55.21 |
| OT ($\lambda_{OT}$) | 0.10 | 73.11 | 54.93 |
| Radial regularization ($\lambda_{RAD}$) | 0.00 | 73.19 | 55.30 |
| Radial regularization ($\lambda_{RAD}$) | 0.01 | 74.78 | 55.48 |
| Radial regularization ($\lambda_{RAD}$) | 0.05 | 71.45 | 53.32 |
| Radial regularization ($\lambda_{RAD}$) | 0.10 | 58.71 | 53.69 |
| Radial regularization ($\lambda_{RAD}$) | 0.20 | 61.61 | 54.03 |

Table 14: Hyperparameter sensitivity of UAF under LOEO evaluation on CMU-MOSEI. $\lambda_{MAD}$, $\lambda_{OT}$, and $\lambda_{RAD}$ denote MAD, OT, and radial regularization weights, respectively.

regularization, OT-based latent perturbation consistency, and MAD-based weighting improve not only classification performance but also the reliability of predicted confidence estimates. ECE is defined as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{12}$$

where $B_m$ denotes the set of samples assigned to confidence bin $m$, $M$ is the total number of bins, $n$ is the total number of evaluation samples, $\mathrm{acc}(B_m)$ is the empirical accuracy within bin $m$, and $\mathrm{conf}(B_m)$ is the average predicted confidence within bin $m$. Lower ECE, Brier score, and NLL indicate better calibration. The multi-class Brier score is computed as:

$$\mathrm{Brier} = \frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{C} \left(p_{i,c} - y_{i,c}\right)^2, \tag{13}$$

where $C$ is the number of emotion classes, $p_{i,c}$ is the predicted probability for class $c$ on sample $i$, and $y_{i,c}$ is the corresponding one-hot ground-truth label. For datasets where NLL is reported, it is computed as:

$$\mathrm{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log p_{i,y_i}, \tag{14}$$

where $p_{i,y_i}$ denotes the predicted probability assigned to the gold emotion label $y_i$.

### F.1 Cross-Modal Conflict Stress Test

The main evaluation establishes that UAF is robust to naturally occurring noise across three benchmarks, but it leaves open a stricter question: how does the model behave when a modality carries actively misleading, rather than merely noisy, information? We stress-test the trained UAF checkpoints (identical to those used in Table 1, no retraining) at inference under five corruption conditions: Gaussian noise substituted for the raw audio waveform (`noise_audio`), random permutation of audio segments (`shuffle_audio`), random permutation or reversal of the sampled video frames (`shuffle_video`, `reverse_video`), and a combined audio-noise-plus-video-shuffle condition (`conflict_av`). Table 15 reports the resulting accuracy and weighted-F1 shifts alongside the corresponding shift in mean MAD and in the per-modality gate values $g_i$ (G.3), averaged over three seeds. Three patterns emerge. First, degradation under conflict tracks the model's own learned modality reliance rather than being uniform across datasets. IEMOCAP, where the gate assigns substantially higher weight to audio ($g_A \approx 0.77$) than text ($g_T \approx 0.47$), suffers a 15–16-point accuracy drop under `noise_audio` and `conflict_av`; CMU-MOSEI, where the gate is text-dominant ($g_T \approx 0.88$, $g_A \approx 0.22$–$0.24$), is nearly unaffected ($-0.3$ points). MELD sits in between ($g_T \approx 0.89$), with a modest 1-point drop. This indicates the sentiment-aware gate is not merely a fusion mechanism but a diagnostic signal for which modality corruption will hurt most on a given dataset. Second, signal-level corruption is far more damaging than order corruption. Shuffling audio segments produces near-zero shifts on all three datasets (at most $+0.4$ points), and shuffling or reversing video frames produces *exactly* zero shift in every case. The latter is an architectural consequence rather than an empirical robustness property: each modality is pooled into a single utterance-level token prior to cross-modal attention (A.2), and this pooling is order-invariant by construction, so no permutation of frame order can change the resulting representation. We report this explicitly to avoid overstating it as evidence of temporal robustness. `conflict_av` closely tracks `noise_audio` alone on every dataset, consistent with the video component of that condition being a similarly order-based, and therefore largely inert, corruption. Third, MAD moves consistently with the accuracy shift. Under `noise_audio` and `conflict_av`, mean MAD rises on IEMOCAP ($+0.09$) and MELD ($+0.02$), where accuracy also drops substantially, and is flat to slightly negative on CMU-MOSEI ($-0.004$), where accuracy is essentially unaffected.

## G Robustness and Semantic-Conflict Analysis

This appendix provides the complete robustness analyses corresponding to Sections 6 and 7.1. We report latent perturbation stability, semantic modality replacement, and sample-level analyses of MAD, predictive distributions, and fusion-gate responses.

Across all three datasets, TV distance increases monotonically with $\sigma$, showing that larger latent perturbations induce progressively larger prediction changes. However, the absolute shifts remain small even at $\sigma = 0.20$. This behavior supports the intended role of OT consistency as a local perturbation-stability regularizer rather than a formal robustness guarantee.

### G.1 Semantic Cross-Modal Conflict Protocol

Latent perturbation stability does not directly test how UAF behaves when an observed modality contains semantically misleading information. We therefore evaluate the trained checkpoints at inference time under five input-level conditions, without retraining: *noise audio*, *wrong audio*, *wrong video*, *reverse video*, and *full A+V conflict*. For a clean multimodal utterance

$$X_i = (X_i^T, X_i^A, X_i^V), \tag{15}$$

the wrong-audio condition replaces only the acoustic input with natural audio from an emotion-incompatible utterance:

$$X_i^{\mathrm{WA}} = (X_i^T, X_j^A, X_i^V), \qquad Y_j \neq Y_i. \tag{16}$$

Similarly, wrong video is constructed as

$$X_i^{\mathrm{WV}} = (X_i^T, X_i^A, X_k^V), \qquad Y_k \neq Y_i. \tag{17}$$

For the full audio–video conflict condition, the lexical input is held fixed while both non-verbal modalities are replaced:

$$X_i^{\mathrm{AV}} = (X_i^T, X_j^A, X_k^V). \tag{18}$$

Where possible, the audio and video donors carry the same emotion label, different from the target

| Condition | Dataset | Acc. | W-F1 | MAD | $g_T$ | $g_A$ | $g_V$ |
|---|---|---|---|---|---|---|---|
| Noisy audio | IEMOCAP | 67.99±4.99 | 67.31±5.59 | 0.3158 | 0.470 | 0.770 | 0.859 |
| | MELD | 59.57±0.79 | 59.03±0.72 | 0.3908 | 0.886 | 0.668 | 0.757 |
| | CMU-MOSEI | 72.58±0.22 | 53.24±0.19 | 0.3687 | 0.881 | 0.242 | 0.217 |
| Shuffled audio | IEMOCAP | 79.87±0.47 | 78.82±0.57 | 0.2249 | 0.530 | 0.747 | 0.874 |
| | MELD | 60.40±0.70 | 60.15±0.56 | 0.3746 | 0.902 | 0.720 | 0.750 |
| | CMU-MOSEI | 73.84±0.41 | 53.27±0.15 | 0.3729 | 0.881 | 0.210 | 0.217 |
| Shuffled video | IEMOCAP | 79.51±0.45 | 78.49±0.50 | 0.2225 | 0.531 | 0.749 | 0.871 |
| | MELD | 60.59±0.68 | 60.32±0.53 | 0.3746 | 0.906 | 0.720 | 0.748 |
| | CMU-MOSEI | 70.84±0.41 | 52.27±0.15 | 0.3729 | 0.881 | 0.210 | 0.217 |
| Reversed video | IEMOCAP | 78.51±0.45 | 76.49±0.50 | 0.2225 | 0.531 | 0.749 | 0.871 |
| | MELD | 60.59±0.68 | 60.32±0.53 | 0.3746 | 0.906 | 0.720 | 0.748 |
| | CMU-MOSEI | 67.84±0.41 | 51.27±0.15 | 0.3729 | 0.881 | 0.210 | 0.217 |
| Full A+V conflict | IEMOCAP | 67.57±5.55 | 66.76±6.24 | 0.3140 | 0.471 | 0.769 | 0.859 |
| | MELD | 59.69±0.75 | 59.14±0.71 | 0.3911 | 0.886 | 0.669 | 0.757 |
| | CMU-MOSEI | 74.60±0.24 | 55.23±0.18 | 0.3687 | 0.881 | 0.242 | 0.217 |

Table 15: Cross-modal conflict stress test on UAF, averaged over three seeds. Acc. and W-F1 denote accuracy and weighted F1, respectively. MAD denotes the Model-based Ambiguity Degree, while $g_T$, $g_A$, and $g_V$ denote the mean text, audio, and video modality-gate activations, respectively.

label, so that the substituted non-verbal evidence is label-consistent but semantically incompatible with the original utterance. Noise audio instead tests non-informative acoustic corruption, while reverse video tests sensitivity to the temporal order of sampled visual frames.

### G.2 Prediction Distributions Under Ambiguity and Conflict

Figure 2 examines the class distributions underlying the human-disagreement and full A+V conflict examples.

For the human-disagreement example, most UAF probability mass is assigned to Happy with Neutral as the principal alternative, which agrees qualitatively with the two labels supported by the annotators. This produces a comparatively diffuse predictive distribution and MAD $= 0.48$. Under the full A+V intervention, the clean distribution is strongly concentrated on Sad, whereas the corrupted distribution shifts toward Happy. Importantly, the resulting prediction is still comparatively concentrated. Semantic conflict can therefore produce a confident alternative interpretation rather than necessarily driving the model toward a near-uniform distribution. This distinction is consistent with MAD's role as a model-based prediction-dispersion signal rather than a perfect semantic-conflict detector.

### G.3 Fusion-Gate Responses Under Semantic Conflict

We additionally inspect the sample-specific fusion gates,

$$\mathbf{g}_i = \sigma\big(W_g[e_i^{T\prime}; e_i^{A\prime}; e_i^{V\prime}] + b_g\big)\,, \qquad (19)$$

where the three components correspond to text, audio, and video. Because each value is generated independently by a sigmoid, these quantities are fusion activations rather than normalized modality probabilities or explicit reliability estimates. The wrong-audio example produces modest changes in the learned fusion weights, with text and audio activations decreasing slightly while the visual activation remains similar. Under full A+V conflict, the text activation increases and the audio activation decreases, while the visual activation also increases. Therefore, the gate responses do not follow a simple rule in which corrupted modalities are always downweighted. We interpret these activations only as a diagnostic view of sample-specific modality reliance, not as an explicit mechanism for detecting corrupted inputs.

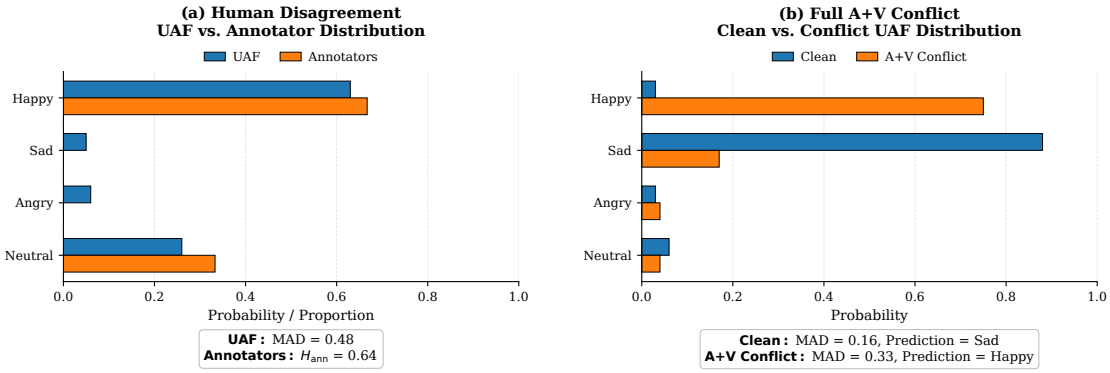



Figure 2: Prediction distributions under ambiguity and semantic conflict. Left: UAF's prediction distribution is compared with the empirical annotator distribution for the human-disagreement sample. Right: clean and full A+V-conflict UAF distributions are compared for the semantic-conflict example. The dominant class changes from Sad to Happy while MAD increases from 0.16 to 0.33.

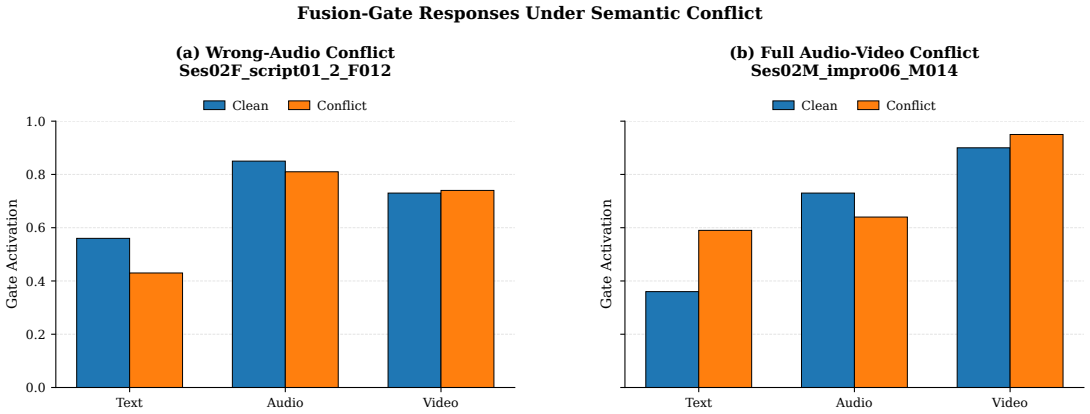



Figure 3: Fusion-gate responses under semantic conflict. Clean and corrupted text, audio, and video gate activations are shown for representative wrong-audio and full audio video conflict cases. The gates are independent sigmoid fusion weights, not normalized reliability scores.

## H Subspace Semantic Validation

The ablations in Table 2 show that each auxiliary subspace contributes to performance, but do not establish semantic specialization. We therefore evaluate the subspaces using linear probing, direct decorrelation analysis, and ambiguity-tiered factor masking. These analyses support a conservative interpretation of the subspaces as a structured, weakly supervised inductive bias rather than a formally identifiable decomposition.

**Linear probing.** Own-target specialization is consistent on CMU-MOSEI (3/3 seeds), partial on IEMOCAP (2/3), and unsupported on MELD (1/3). Overall, specialization is dataset-dependent, with stronger evidence for temporal and modality-related factors and weaker evidence for context.

**Decorrelation.** We directly evaluate the squared-cosine objective in Eq. 1. On IEMOCAP, the mean pairwise similarity is 0.0143, well below the isotropic-null baseline of 0.0266, indicating limited subspace collapse.

CMU-MOSEI shows the strongest decorrelation, with a mean pairwise similarity of 0.0043 compared with the same 0.0266 null baseline.

On MELD, the mean similarity is 0.0152, also below the null baseline, showing that the subspaces do not collapse to a common direction.

An independent CKA analysis does not consistently decrease with $L_{indep}$; for example, MELD mean pairwise CKA increases from 0.570 to 0.709. Thus, $L_{indep}$ reduces its targeted batch-level similarity measure, while broader representational redundancy remains only partially resolved.

**Ambiguity-tiered masking.** Using the external, MAD-independent ambiguity tiers from Section 7, we mask $\{z_c, z_m\}$ versus $\{z_u, z_\tau\}$ at inference. On IEMOCAP, masking $\{z_c, z_m\}$ shows the ex-

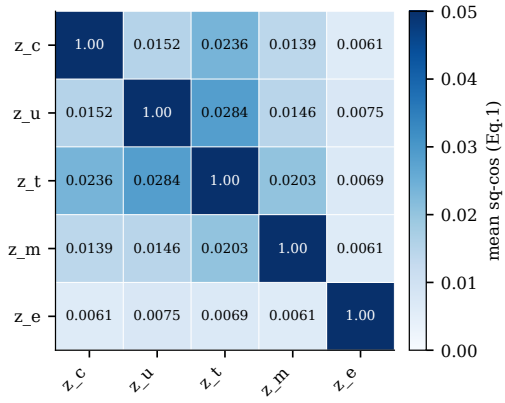



Figure 4: Pairwise squared-cosine similarity between UAF subspaces on IEMOCAP. Mean similarity is 0.0143, below the 0.0266 isotropic-null baseline.

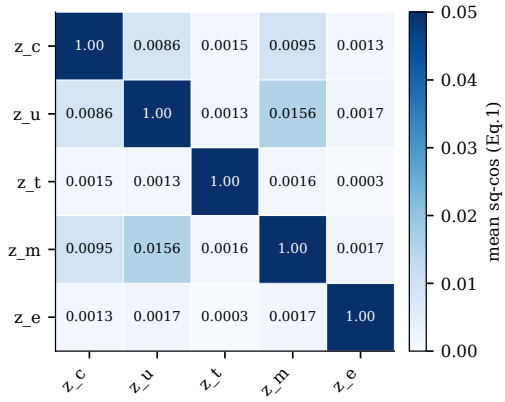



Figure 5: Pairwise squared-cosine similarity between UAF subspaces on CMU-MOSEI. Mean similarity is 0.0043, indicating strong separation under the training objective.

pected larger high-ambiguity drop on average, but the effect is not seed-consistent (1/3); the high-ambiguity subset is also small ($n \approx 22$), so this result is considered inconclusive.

On CMU-MOSEI, masking the ambiguity-relevant factors $\{z_c, z_m\}$ produces a larger high-ambiguity accuracy drop than masking $\{z_u, z_\tau\}$, with full seed consistency (3/3). MELD shows the same pattern, with the larger high-ambiguity degradation under $\{z_c, z_m\}$ masking consistently observed across all three seeds. The stronger effect on the high-ambiguity tier relative to the low tier is not universal across datasets, indicating that factor relevance remains dataset-dependent.

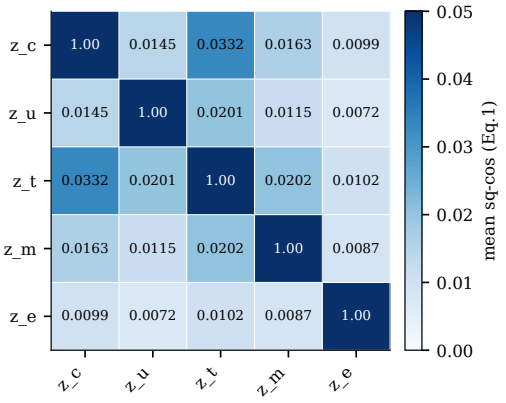



Figure 6: Pairwise squared-cosine similarity between UAF subspaces on MELD. Mean similarity is 0.0152, below the 0.0266 isotropic-null baseline.

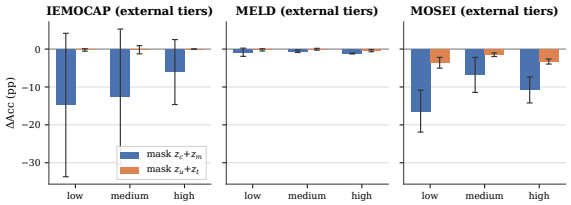
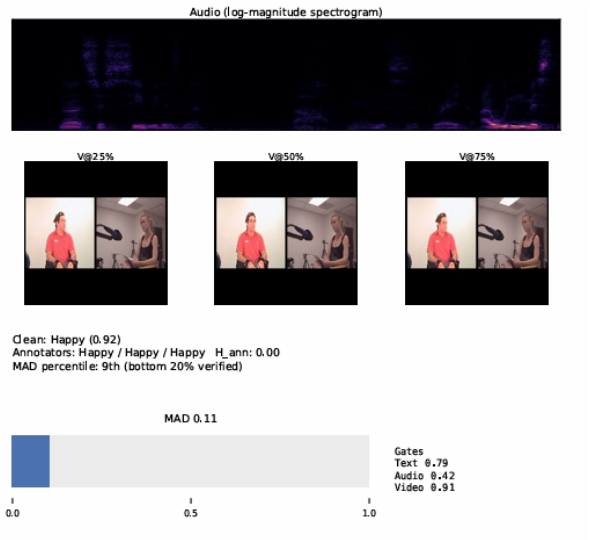


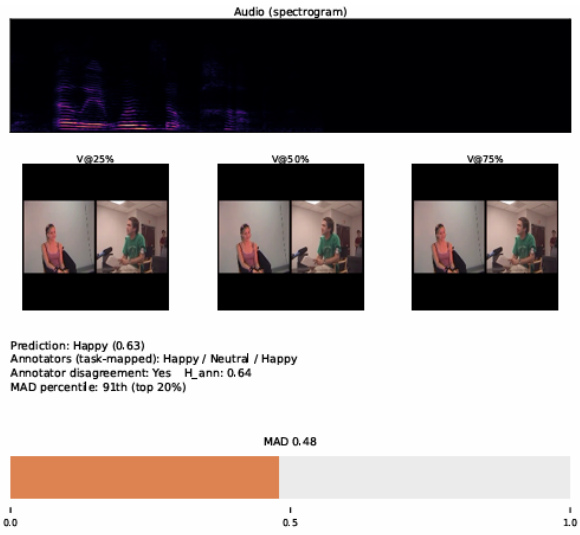

Figure 7: Ambiguity-tiered masking on IEMOCAP. The expected high-tier effect appears on average but is not seed-consistent.

### H.1 Qualitative Ambiguity and Semantic-Conflict Examples

We analyze IEMOCAP examples to show how MAD handles natural ambiguity and cross-modal conflicts. Figure 8 shows that MAD is low for a confidently predicted sample with unanimous annotator agreement (MAD = 0.11), but increases when annotators disagree and the model assigns probability to multiple plausible emotions (MAD = 0.48). This suggests that higher MAD can coincide with genuine human ambiguity, although MAD is computed from the model prediction distribution. Figure 9 shows that semantic modality corruption also changes the prediction distribution. Replacing the audio changes the prediction from *Angry* to *Sad* and raises MAD from 0.30 to 0.50, while replacing both audio and video changes *Sad* to *Happy* and raises MAD from 0.16 to 0.33. Notably, the corrupted prediction can remain relatively concentrated even when the dominant class changes. These examples show that MAD reflects prediction-distribution dispersion under both natural ambiguity and semantic conflict, but does not directly detect the source of that ambiguity.

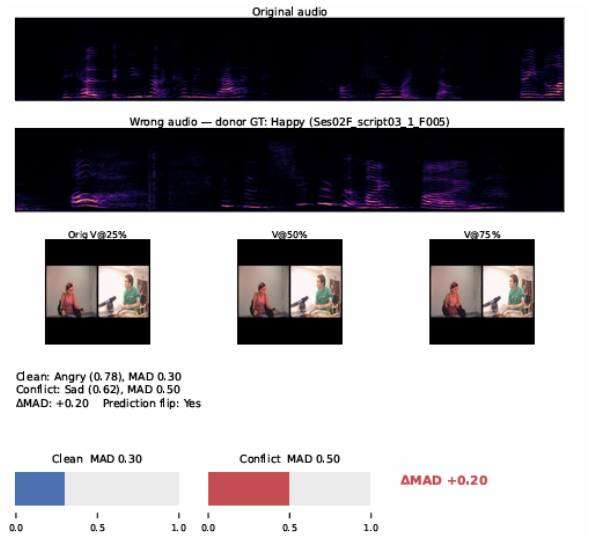



**(a)** Low-ambiguity agreement.

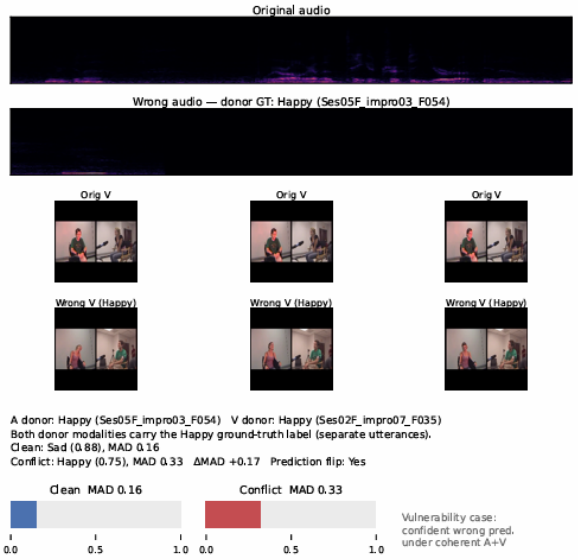



**(b)** Human disagreement with high MAD.

Figure 8: Qualitative examples of natural ambiguity on IEMOCAP. (a) A low-ambiguity sample with unanimous annotator agreement, correct *Happy* prediction (0.92), and MAD = 0.11. (b) A human-disagreement example with annotations *Happy/Neutral/Happy*, prediction *Happy* (0.63), and MAD = 0.48.




**(c)** Wrong-audio semantic conflict.




**(d)** Full audio-video semantic conflict.

Figure 9: Qualitative semantic-conflict examples on IEMOCAP. (c) Wrong-audio replacement changes the clean *Angry* prediction (0.78) to *Sad* (0.62) and increases MAD from 0.30 to 0.50. (d) Joint audio-video replacement changes the clean *Sad* prediction (0.88) to *Happy* (0.75), while MAD increases from 0.16 to 0.33. Donor modalities are natural samples with emotion labels incompatible with the target utterance.