CREL: Causal Retrieval Evidential Learning for Robust Multimodal Sentiment Analysis with Missing Modalities

Anonymous ACL submission

Abstract

Multimodal Sentiment Analysis (MSA) in real-world deployment often encounters modality missingness due to sensor failures, privacy masking, or transmission loss. Most existing methods complete missing semantics through continuous completion or robustness reweighting within a closed world correlation model, which tends to produce mean regression artifacts under one-to-many affective mappings and to output overconfident softmax point estimates when evidence is scarce. To address these issues, we propose **Causal Retrieval Evidential Learning (CREL)**, which reframes modality completion as evidence reasoning by employing Retrieval Augmented Interaction (RAI) to query discrete semantic evidence from a Prototype-based Multimodal Knowledge Base (MKB), purifying the retrieved cues against spurious correlations via Causal De-confounding Adjustment (CDA), and explicitly modeling prediction confidence through Evidential Uncertainty Calibration (EUC) grounded in subjective logic to reflect evidence sufficiency. Extensive experiments on benchmark datasets show that CREL achieves SOTA performance under missing modalities, and ablation studies further highlight the contribution of each designed component.

1 Introduction

In real-world settings, accurate sentiment understanding relies on integrating heterogeneous cues from language, acoustic, and visual (Tsai et al., 2019; Rahman et al., 2020; Meng et al., 2025). However, deployed systems frequently suffer from modality missingness due to sensor failures, privacy masking, or transmission loss. Such missingness can occur as Intra-Modal Missingness (IMM), where portions of a modality sequence are absent or corrupted, or as Fixed-Modal Missingness (FMM), where one or more entire modalities are unavailable (Li et al., 2025; Zhu et al., 2025). Because each modality provides complementary and partially non-redundant affective evidence, missing information weakens cross-modal corroboration and increases ambiguity, which in turn degrades sentiment prediction performance (Yang et al., 2025; Yuan et al., 2023).

Existing work can be broadly categorized into two families. Category A focuses on continuous completion, where missing representations are inferred through continuous mappings, including reconstruction and imputation (Yuan et al., 2021, 2023; Yang et al., 2025), as well as proxy or distillation variants that regress cross-modal surrogates (Zhu et al., 2025; Li et al., 2024; Sun et al., 2024b). Category B emphasizes robust fusion and adaptation, aiming to reduce sensitivity to missing channels by learning modality-invariant and modality-specific factors (Hazarika et al., 2020), adopting self-adaptive fusion under uncertain missingness (Li et al., 2025; Gao et al., 2024), applying gradient modulation and reweighting for robustness (Zhou et al., 2024; Guo et al., 2024), or adapting unimodal models and prompting schemes to operate with absent modalities (Li et al., 2023a; Zhang et al., 2024b,a). Despite these differences in methodology, both families share the same goal of enabling reliable prediction from incomplete multimodal evidence.

These families exhibit two shared failure modes that become dominant under severe modality missingness. First, mean regression artifacts arise because affective semantics are one-to-many: the same text can convey sarcasm, anger, or surprise depending on prosody and gesture. When completion is optimized under a regression style risk, explicitly through ℓ_1 or ℓ_2 reconstruction or implicitly through proxy regression, ambiguity drives the solution toward conditional means or medians (Yang et al., 2025; Zhu et al., 2025; Li et al., 2024). As a result, the completed representations become overly smooth and class agnostic, atten-

uating fine grained cues that are critical for sentiment and emotion prediction (Tsai et al., 2019; Zadeh et al., 2017). In real world decision settings, this smoothing can mask rare but decisive affective signals, yielding incorrect polarity or intensity estimates that are difficult to audit. Second, deterministic overconfidence under evidence scarcity arises because most pipelines ultimately rely on softmax point estimates that normalize probabilities regardless of how much evidence supports the prediction (Guo et al., 2017). As missingness increases, this constraint can produce low entropy predictions even when inputs provide little valid evidence, so the model cannot express epistemic ignorance (Guo et al., 2017; Lakshminarayanan et al., 2017). In practice, this prevents the system from warning users about unreliable inputs and can lead to silent failures in human facing applications.

To address these issues, we propose **Causal Retrieval Evidential Learning (CREL)**, an evidence centric paradigm that replaces feature hallucination with evidence reasoning. In particular, CREL constructs a prototype based **Multimodal Knowledge Base (MKB)** and completes missing modalities by retrieving discrete evidence through a Retrieval **Augmented Interaction (RAI)** module, rather than regressing continuous features. By inductively distilling unimodal prototypes as keys and their fused semantic centroids as values, so completion is grounded in denoised discrete evidence. To reduce confounder driven shortcuts in retrieval, CREL introduces a **Causal De-confounding Adjustment (CDA)** objective that penalizes retrieval representations aligned with nuisance factors, such as speaker identity or channel artifacts, via counterfactual contrastive regularization. To prevent overconfident decisions under evidence scarcity, CREL performs **Evidential Uncertainty Calibration (EUC)** under Subjective Logic by mapping aggregated evidence to a Dirichlet distribution whose epistemic uncertainty increases when evidence is insufficient. The main contributions of our paper are as follows:

- We propose **CREL**, which treats modality completion as evidence reasoning rather than continuous reconstruction to mitigate mean regression artifacts and reduce overconfident predictions.
- We build an inductively constructed prototype memory **MKB** together with a **RAI** module that queries prototypes and fuses their retrieved fused semantic values with observed representations, supporting discrete evidence completion and improving completion fidelity under one-to-many affective mappings.
- We introduce a deconfounding contrastive objective (**CDA**) to suppress confounder driven retrieval and integrate evidential learning under Subjective Logic (**EUC**) to quantify epistemic uncertainty, enabling predictions whose confidence better tracks evidence sufficiency.

Related Work can be found in Appendix A.

2 Methodology

We propose CREL, a framework designed to address the twin challenges of mean-regression artifacts and deterministic overconfidence in multimodal sentiment analysis under modality missingness. CREL fundamentally reframes modality completion from a closed-world generation task to evidence reasoning process. As illustrated in Figure 1, the architecture comprises four synergistically coupled components. A prototype-based MKB provides discrete evidence units constructed from training data. RAI retrieves and integrates evidence conditioned on observed modalities and estimates an evidence reliability score. CDA regularizes the retrieval representation used by RAI to suppress confounder-driven shortcuts. EUC maps the fused representation into a Dirichlet predictive distribution whose vacuity increases when evidence is weak.

2.1 Problem Formulation

Let $\mathcal{U} = \{l, a, v\}$ denote the language, acoustic, and visual modalities. The dataset is defined as $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, where $X_i = \{x_i^m\}_{m\in\mathcal{U}}$ represents the multimodal input and y_i is the label for sentiment classes. The missingness status is formalized by an availability mask $\mathbf{s}_i \in \{0,1\}^{|\mathcal{U}|}$, defining the observed set $\Omega_i = \{m \in \mathcal{U} \mid s_{i,m} = 1\}$. For each observed modality $m \in \Omega_i$, a unimodal encoder f_m extracts a feature vector $h_i^m = f_m(x_i^m) \in \mathbb{R}^d$, where d is the unified feature dimension. To form a query basis, we aggregate observed features into a summary vector $h_{i,\Omega}$ using a permutation-invariant aggregation function $\mathrm{Agg}(\cdot)$ (implemented via attention pooling). Our objective is to accurately predict y_i using only $X_{i,\Omega}$ while simultaneously maximizing the epistemic uncertainty estimate T_i when the provided evidence

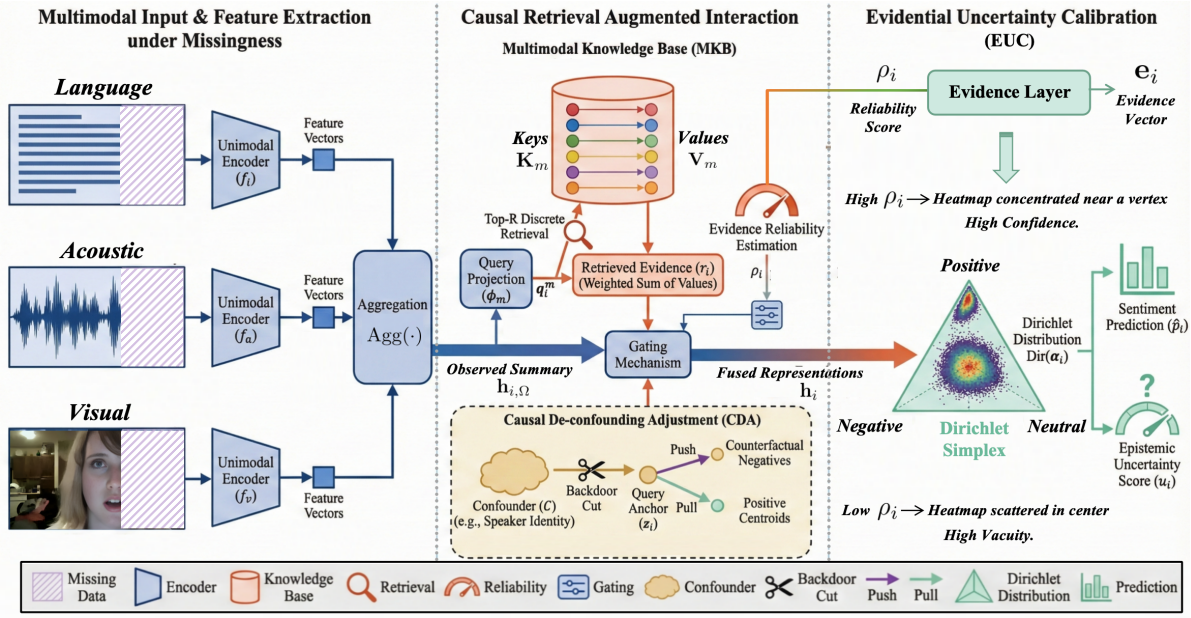

Figure 1: Overview of the Causal Retrieval Evidential Learning (CREL) framework. CREL mitigates modality missingness via evidence reasoning rather than continuous regression. It leverages Retrieval Augmented Interaction (RAI) to query discrete prototypes from a memory bank MKB, employs CDA to purify retrieved cues against spurious correlations, and uses EUC to produce reliable predictions with quantified epistemic uncertainty.

is insufficient to support a reliable decision.

$$h_{i,\Omega} = \mathrm{Agg}\left(\{h_i^m \mid m \in \Omega_i\}\right) \in \mathbb{R}^d. \quad (1)$$

2.2 Multimodal Knowledge Base (MKB)

To fundamentally mitigate the mean-regression artifacts caused by continuous reconstruction of one-to-many affective mappings (as discussed in Sec. 1), we propose to replace feature generation with discrete evidence retrieval. Consequently, we design an inductively constructed MKB that stores crystallized semantic prototypes rather than raw instances. To prevent information leakage, MKB is built strictly from the training split $\mathcal{D}_{\text{train}}$. For each modality $m \in \mathcal{U}$, we perform K-means clustering on all unimodal training features to extract P centroids, serving as the discrete prototype keys:

$$K_m = \{\mu_p^m\}_{p=1}^P, \qquad \mu_p^m \in \mathbb{R}^d. \quad (2)$$

These keys represent the canonical patterns of each modality (e.g., a specific facial expression cluster). To capture the cross-modal semantic completion corresponding to these patterns, we define a fusion extractor $g(\cdot)$ to compute the joint representation $u_i = g(X_i)$ for complete training samples. For each prototype μ_p^m, we identify its assigned training sample set $\mathcal{I}_p^m$ and compute the Fused Semantic Value v_p^m:

$$v_p^m = \frac{1}{|\mathcal{I}_p^m|} \sum_{i \in \mathcal{I}_p^m} u_i \in \mathbb{R}^d. \quad (3)$$

The final memory bank $\mathcal{M} = \{(K_m, V_m)\}_{m \in \mathcal{U}}$ maps unimodal cues to complete multimodal semantics. This design ensures that retrieved evidence v_p^m possesses the high-frequency variance of real-world data, effectively avoiding the smoothing effect inherent in regression-based methods.

2.3 Retrieval Augmented Interaction (RAI)

RAI is designed to execute the evidence reasoning by dynamically querying the MKB. Unlike methods that hallucinate missing features from internal correlations, RAI retrieves external evidence to fill the semantic gap. First, we project the observed summary $h_{i,\Omega}$ into the query space of each modality m using a learnable projection $\phi_m : \mathbb{R}^d \to \mathbb{R}^d$:

$$q_i^m = \phi_m(h_{i,\Omega}) \in \mathbb{R}^d. \quad (4)$$

We then compute the cosine similarity between q_i^m and keys in K_m, retrieving the Top-R nearest prototypes to form the index set $\mathcal{J}_{i,m}$. The modality-specific evidence $r_{i,m}$ is obtained by aggregating the values V_m weighted by the softmax-normalized similarities. These are then pooled to form the global retrieved evidence r_i.

Evidence Reliability Estimation. A critical innovation of CREL is the recognition that retrieval is not always reliable, naively integrating irrelevant prototypes introduces noise. To quantify this, we calculate the Evidence Reliability Score $\rho_{i,m}$ based on the entropy of the retrieval distribution weights $\mathbf{w}_i^m$:

$$\rho_{i,m} = 1 - \frac{H(\mathbf{w}_i^m)}{\log R} \in [0, 1] \tag{5}$$

where $H(\mathbf{w}) = -\sum_{j=1}^{R} w_j \log(w_j + \epsilon)$. A low entropy (high $\rho_{i,m}$) implies the query strongly matches specific prototypes, indicating reliable evidence. Conversely, a uniform distribution implies ambiguity. The sample-level reliability is $\rho_i = \text{Mean}(\{\rho_{i,m}\}_{m \in \Omega_i})$. Finally, we employ a reliability-aware gating mechanism γ_i to fuse the observed and retrieved information:

$$h_i = \gamma_i \odot h_{i,\Omega} + (1 - \gamma_i) \odot r_i. \tag{6}$$

Here, $\gamma_i = \sigma\big(W_\gamma[h_{i,\Omega}; r_i; \rho_i]\big)$, ρ_i explicitly informs the gate to down-weight r_i when retrieval is uncertain, ensuring robustness.

2.4 Causal De-confounding Adjustment

To address the selection bias issue where retrieval is dominated by nuisance confounders (e.g., background noise) rather than semantic content, we introduce the CDA module. We posit a Structural Causal Model where the confounder C creates a spurious backdoor path between the query Q and the retrieved evidence R. To estimate the true causal effect, we employ a counterfactual contrastive learning objective.

Counterfactual Sampling. We generate a set of hard negative samples that are confounder-aligned but label-mismatched. Specifically, we utilize a separate lightweight encoder $\psi(\cdot)$ to extract superficial features (representing confounders) $\hat{c}_i$. For an anchor i, we select a counterfactual negative set $\mathcal{N}_i^{\text{cf}}$ consisting of samples j that have high similarity to $\hat{c}_i$ but possess different sentiment labels $y_j \neq y_i$.

De-confounding Loss. We enforce the retrieval representation z_i ($z_i = \text{Pool}\left(\{q_i^m \mid m \in \Omega_i\}\right) \in \mathbb{R}^d$) to be distinct from these counterfactual negatives. The loss function is defined as:

$$\mathcal{L}_{\text{CDA}} = \sum_{i \in \mathcal{B}} \sum_{n \in \mathcal{N}_i^{\text{cf}}} \text{softplus}\Big(m_0 + \text{sim}(z_i, z_{neg}) - \text{sim}(z_i, \bar{z}_i^+)\Big) \tag{7}$$

where $\bar{z}_i^+$ is the centroid of positive samples and m_0 is a margin. z_{neg} denotes the retrieval representation of a counterfactual negative $n \in \mathcal{N}_i^{\text{cf}}$. By penalizing high similarity with $\mathcal{N}_i^{\text{cf}}$, CDA effectively forces the query encoder ϕ_m to discard nuisance information (like speaker identity) and focus more on affective semantics, ensuring the validity of the retrieved evidence.

2.5 Evidential Uncertainty Calibration (EUC)

To resolve the deterministic overconfidence problem inherent in softmax-based classifiers, EUC grounds the prediction in Subjective Logic. We model the sentiment prediction not as a point estimate, but as a Dirichlet distribution parameterized by evidence counts, allowing the model to express "I don't know" through the vacuity dimension.

Evidence Formulation. The core novelty lies in how we construct the evidence vector $\mathbf{e}_i$. Unlike standard evidential learning, we map the fused representation h_i to nonnegative evidence via softplus and modulate it by the evidence reliability score $\rho_i \in [0, 1]$ from RAI:

$$\mathbf{e}_i = \rho_i \cdot \text{softplus}(W_e h_i) \in \mathbb{R}^C_{\geq 0}, \tag{8}$$

This modulation ensures that if the retrieval quality is poor ($\rho_i \to 0$), the total evidence decreases, preventing the model from making confident predictions based on noisy completion.

Uncertainty Quantification. We then set the Dirichlet parameters as $\boldsymbol{\alpha}_i = \mathbf{e}_i + \mathbf{1}$, where adding $\mathbf{1}$ corresponds to a non-informative unit prior and ensures positivity. The Dirichlet strength $S_i = \sum_{c=1}^{C} \alpha_{i,c}$, measures the total amount of evidence and controls distribution sharpness.

To quantify epistemic uncertainty due to evidence scarcity, we output the vacuity score $u_i = \frac{C}{S_i}$. Under severe missingness, if RAI fails to find reliable evidence, ρ_i drops, causing $\mathbf{e}_i \to \mathbf{0}$, $S_i \to C$, and $u_i \to 1$. This encourages the model to exhibit high uncertainty in low-information regimes.

2.6 Training Objective

CREL is trained with a joint objective that couples the three modules. CDA regularizes the retrieval representation z_i to suppress confounder-driven shortcuts, ensuring semantically valid evidence. RAI then fuses the resulting evidence into h_i, and EUC maps (h_i, ρ_i) to a Dirichlet predictive distribution. Accordingly, the composite loss

jointly reduces evidential Bayes risk and deconfounding bias.

Evidential Risk Minimization. Given the Dirichlet parameters $\boldsymbol{\alpha}_i$, we minimize the expected Sum-of-Squares error between the one-hot label y_i and the predictive distribution. The loss analytically decomposes into prediction error, variance penalty, and regularization terms:

$$\mathcal{L}_{\mathrm{EDL}} = \sum_{i=1}^{N} \Bigg[\underbrace{\sum_{c=1}^{C} (y_{i,c} - \hat{p}_{i,c})^2}_{\text{Prediction Error}} + \underbrace{\sum_{c=1}^{C} \frac{\hat{p}_{i,c}(1-\hat{p}_{i,c})}{S_i + 1}}_{\text{Variance Penalty}} \Bigg] + \lambda_{\mathrm{KL}} \sum_{i=1}^{N} \mathrm{KL}\big(\mathrm{Dir}(\boldsymbol{\alpha}_i) \,\|\, \mathrm{Dir}(\mathbf{1})\big). \tag{9}$$

Here, $\hat{p}_{i,c} = \alpha_{i,c}/S_i$. The first two terms jointly encourage accuracy and evidence accumulation (increasing S_i). The KL divergence term regularizes the distribution toward the uniform prior $\mathrm{Dir}(\mathbf{1})$, the state of maximum epistemic uncertainty, to prevent overconfidence in low-evidence regimes. λ_{KL} is an annealing coefficient that gradually penalizes unjustified confidence during training.

Total Objective. The final objective combines the evidential risk with the causal de-confounding constraint proposed in Sec. 2.4:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{EDL}} + \lambda_{\mathrm{CDA}} \mathcal{L}_{\mathrm{CDA}}, \tag{10}$$

where $\lambda_{\mathrm{CDA}} > 0$ is a hyperparameter balancing the retrieval regularization.

3 Experiments

3.1 Experimental Setup

Benchmarks and Metrics. We evaluate on three public multimodal sentiment benchmarks, CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020). We use the standard data splits and preprocessing. Details are provided in Appendix B. For CMU-MOSI and CMU-MOSEI, we report Acc-2, Acc-5, Acc-7, F1, MAE, and Corr, following the common protocol in prior work (Zhu et al., 2025). For CH-SIMS, we report Acc-2, Acc-3, Acc-5, F1, MAE, and Corr.
Implementation Details. All experiments are conducted in PyTorch on one NVIDIA A100 GPU with 80GB memory. We use BERT as the language encoder. We adopt the same acoustic and visual feature extraction pipeline as the compared methods to ensure comparability (Li et al., 2025; Zhu et al., 2025). Each reported number is the mean over 5 random seeds. Full hyperparameters and baseline descriptions are provided in Appendix C and Appendix D.

3.2 Missingness Settings and Evaluation Protocol

We consider two missingness settings that are widely used in robust multimodal learning.

Intra-Modal Missingness (IMM). IMM block missing a ratio r within each modality stream. We evaluate $r \in \{0.0, 0.1, \ldots, 0.9\}$. For acoustic and visual streams, we replace missing feature vectors with zeros. For language, we replace missing tokens with `[UNK]` and apply the corresponding attention mask. We report robustness as the mean performance averaged over all tested rates. For methods that support missing inputs during training, we apply the same IMM process to the training and test splits for each rate r. For methods that require complete inputs, we train on fully observed training data and evaluate under IMM only at test time. This protocol follows prior (Zhu et al., 2025).
Fixed-Modality Missingness (FMM). FMM removes one or multiple modalities entirely at inference time. All methods are trained on the fully observed training split. At test time, we drop the specified modality streams. For baselines that do not natively support missing inputs, we use zero padding for missing acoustic and visual streams and `[UNK]` for missing language tokens. We report the result for each modality subset and their average.

3.3 Results under Intra-Modal Missingness

Tables 1 and 2 summarize results under IMM on CMU-MOSI, CMU-MOSEI, and CH-SIMS. CREL achieves the best overall performance across classification and regression metrics.

On CMU-MOSI and CMU-MOSEI, Table 1 shows consistent improvements. For example, CREL attains higher Acc-2 under both binary settings and yields the lowest MAE and the highest Corr among all listed methods. On CH-SIMS, CREL improves Acc-2 from 73.64% to 74.96% and improves F1 from 74.65% to 80.07% when compared with the strongest baseline P-RMF. These gains indicate improved robustness under random intra-modal missingness. We attribute the gains to three components that are validated by ablations in Section 3.5. First, RAI retrieves discrete semantic evidence from the prototype based memory

Model	CMU-MOSI						CMU-MOSEI					
	Acc-2	F1	Acc-5	Acc-7	MAE	Corr	Acc-2	F1	Acc-5	Acc-7	MAE	Corr
MISA	70.33/71.49	70.00/71.28	33.08	29.85	1.085	0.524	75.82/71.27	68.73/63.85	39.39	40.84	0.780	0.503
Self-MM	69.26/70.51	67.54/66.60	34.67	29.55	1.070	0.512	77.42/73.89	72.31/68.92	45.38	44.70	0.695	0.498
MMIM	67.06/69.14	64.04/66.65	33.77	31.30	1.077	0.507	75.89/73.32	70.32/68.72	41.74	40.75	0.739	0.489
CENET	67.73/71.46	64.85/68.41	37.25	30.38	1.080	0.504	77.34/74.67	74.08/70.68	47.83	47.18	0.685	0.535
TFR-Net	66.35/68.15	60.06/61.73	34.67	29.54	1.200	0.459	77.23/73.62	71.99/68.80	34.67	46.83	0.697	0.489
ALMT	68.39/70.40	71.80/72.57	33.42	30.30	1.083	0.498	77.54/76.64	78.03/77.14	41.64	40.92	0.674	0.481
LNLN	70.94/72.55	71.25/72.73	38.27	34.26	1.046	0.527	78.19/76.30	79.95/77.77	46.17	45.42	0.692	0.530
P-RMF	71.53/72.81	71.69/72.93	38.50	34.19	1.038	0.525	78.83/78.14	80.39/79.33	45.87	44.63	0.658	0.589
DRA	71.60/73.18	71.51/73.15	38.65	34.47	1.069	0.520	77.48/78.14	77.44/77.51	48.01	47.02	0.666	0.583
CREL	**73.17/74.26**	**73.31/74.39**	**39.87**	**35.79**	**0.986**	**0.552**	**79.38/80.07**	**80.27/80.39**	**48.87**	**48.66**	**0.642**	**0.606**

Table 1: Result on MOSI and MOSEI under Intra-Modal Missingness (IMM). Results are averaged over $r \in \{0.0, 0.1, \ldots, 0.9\}$. a/b represents binary accuracy under negative vs. non-negative and negative vs. positive settings. We further report the mean ± standard deviation calculated over 5 different seeds in Appendix G.

bank. This reduces the tendency of regression style completion to collapse to mean predictions under ambiguous supervision. Second, CDA reduces confounder driven retrieval shortcuts. This becomes important when missingness amplifies spurious correlations. Third, EUC models predictive uncertainty with a Dirichlet distribution and uses retrieval reliability to modulate evidence strength. This reduces overconfident predictions when evidence is scarce.

Figure 2 further reports robustness curves under increasing missing rates. All methods degrade as r increases. CREL shows a consistently stronger trend across most rates, which matches the averaged results in Table 2. More detailed test results are provided in Appendix F.

Model	Acc-2	F1	Acc-3	Acc-5	MAE (↓)	Corr
MISA	72.71	66.30	56.87	31.53	0.539	0.348
Self-MM	72.81	68.43	56.75	32.28	0.508	0.376
MMIM	69.86	66.21	52.76	31.81	0.544	0.339
CENET	68.13	57.90	53.17	22.29	0.589	0.107
TFR-Net	68.13	58.70	52.89	26.52	0.661	0.169
ALMT	71.85	76.21	56.47	34.16	0.509	0.372
LNLN	72.73	79.43	57.14	34.64	0.514	0.397
P-RMF	73.64	74.65	54.75	34.83	0.500	0.414
CREL (Ours)	**74.96**	**80.07**	**57.79**	**36.29**	**0.485**	**0.437**

Table 2: Result on CH-SIMS under Intra-Modal Missingness. Results are averaged over $r \in \{0.0, \ldots, 0.9\}$.

3.4 Results under Fixed-Modality Missingness

We evaluate FMM on CMU-MOSI and CMU-MOSEI, where one or multiple modalities are unavailable at inference. Table 3 reports F1 scores for all modality subsets. CREL achieves the best average performance across the seven test conditions. On CMU-MOSEI, CREL reaches an average F1

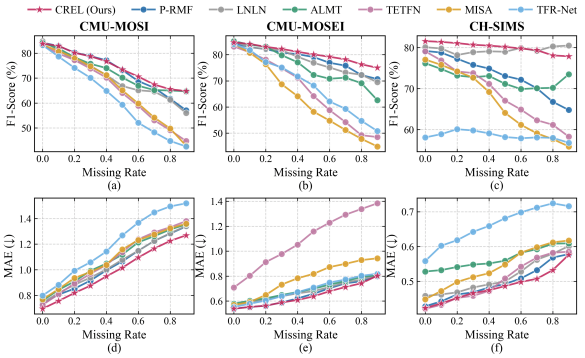

Figure 2: Performance curves of various missing rates.

Models	Test Conditions (F1-score, %)							
	$\{l, v, a\}$	$\{l, v\}$	$\{l, a\}$	$\{v, a\}$	$\{l\}$	$\{v\}$	$\{a\}$	**Average**
CMU-MOSI								
Self-MM	84.64	74.97	69.81	47.12	67.80	38.52	40.95	60.54
DCCAE	66.76	65.43	62.25	42.39	61.78	41.45	41.30	54.48
DMD	84.50	68.45	70.51	50.47	68.97	42.26	43.33	61.21
MMIN	85.20	84.76	83.50	45.51	82.39	43.86	44.25	67.07
NIAT	83.16	81.75	82.03	43.58	80.72	42.79	42.10	65.16
GCNet	83.20	83.58	84.73	70.02	81.12	59.67	68.07	75.77
CIF-MMIN	84.27	82.48	81.62	54.70	81.16	51.09	52.02	69.62
IMDer	84.71	81.67	82.53	57.30	81.40	56.71	55.29	71.37
EMT	82.43	83.30	82.64	55.90	80.37	53.63	51.50	69.97
GGMD	85.70	84.39	83.53	54.60	81.74	52.88	51.24	70.58
CorrKD	83.94	82.41	82.36	73.74	81.20	60.72	66.52	75.84
P-RMF	84.37	81.94	82.10	73.11	81.36	70.32	71.44	77.81
ROSA	86.30	85.39	85.13	83.81	83.64	56.79	81.91	80.42
CREL(Ours)	**87.28**	**85.89**	**85.76**	**84.87**	**84.27**	**70.97**	**82.06**	**83.01**
CMU-MOSEI								
Self-MM	83.69	74.62	75.91	49.52	71.53	37.61	43.57	62.35
DCCAE	75.70	73.16	73.82	45.03	71.19	43.84	41.47	60.60
DMD	84.78	72.45	74.78	52.70	70.26	39.84	46.18	63.00
MMIN	85.78	83.55	82.93	49.35	82.21	47.63	46.08	68.22
NIAT	86.35	85.31	85.10	48.57	83.29	45.76	44.43	68.40
GCNet	82.35	81.15	81.96	69.21	80.52	61.83	66.54	74.79
CIF-MMIN	85.33	83.36	82.70	53.69	81.94	50.70	51.38	69.87
IMDer	84.71	82.39	83.41	53.30	82.77	51.83	50.47	69.84
EMT	85.17	84.45	83.81	52.39	83.04	50.79	51.28	70.13
GGMD	87.07	85.52	83.10	52.67	81.45	49.96	50.39	70.02
CorrKD	82.16	81.28	81.74	71.92	80.76	62.30	66.09	75.18
P-RMF	85.48	85.17	84.61	76.88	81.91	73.19	75.91	80.45
ROSA	89.56	87.32	86.64	84.19	84.42	54.38	79.04	80.79
CREL(Ours)	**90.66**	**88.16**	**87.58**	**84.72**	**84.56**	**73.57**	**79.82**	**84.15**

Table 3: Result on the MOSI, MOSEI With FMM.

of 84.15%, which exceeds the strongest baseline ROSA at 80.79%. CREL also performs well in low evidence subsets. Under the visual only setting $\{v\}$, CREL attains 73.57% and slightly improves upon P-RMF at 73.19%. Under the audio only setting $\{a\}$, CREL attains 79.82% and improves upon ROSA at 79.04%. On CMU-MOSI, CREL achieves the best average F1 of 83.01%.

The same evidence centric principle used in IMM remains effective in FMM. RAI compensates missing modalities through prototype level evidence retrieval. CDA reduces shortcut retrieval when observed modalities are limited. EUC avoids deterministic overconfidence in low evidence subsets by explicitly modeling uncertainty.

Setting	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	F1(%)	MAE ↓	F1(%)	MAE ↓	F1(%)	MAE ↓
CREL	**74.39**	**0.986**	**80.39**	**0.642**	**80.07**	**0.485**
w/o MKB^1	71.42	1.020	77.94	0.673	76.81	0.512
w/o RAI	68.36	1.086	75.22	0.725	73.56	0.562
w/o RAI^2	70.28	1.043	77.11	0.688	76.19	0.526
w/o CDA	69.17	1.066	76.28	0.709	75.29	0.544
w/o EUC	73.06	1.000	79.03	0.655	78.30	0.507
w/o RAF.	72.92	1.011	78.89	0.664	77.86	0.502

Table 4: Ablation studies of CREL with IMM.

3.5 Ablation studies

We conduct ablations under the same IMM protocol. Each variant changes one factor while keeping the backbone and optimization unchanged. Table 4 reports results.

Effect of evidence retrieval. Removing retrieval (w/o RAI) causes the largest degradation on all three datasets. This indicates that external evidence is essential when intra-modal missingness weakens cross modal corroboration. Replacing Top-R prototype retrieval with a Regression based Imputation Module (w/o RAI^2) also underperforms CREL. This supports the use of discrete evidence units under ambiguous supervision.

Effect of memory abstraction. Replacing the prototype based memory bank with instance memory (w/o MKB^1) reduces performance on all datasets. This suggests that prototype level abstraction provides cleaner and more stable evidence units under varying missing rates.

Effect of De-confounding and reliability control. Removing CDA (w/o CDA) yields a clear drop, which supports the need to suppress confounder driven retrieval shortcuts. Removing Reliability-Aware Fusion (w/o RAF.) also degrades performance. This implies that retrieved cues should be down weighted when retrieval is uncertain.

Effect of evidential calibration. Removing EUC (w/o EUC) consistently harms performance and increases MAE, especially on CH-SIMS. This indicates that evidential learning improves confidence behavior and mitigates error accumulation under scarce evidence.

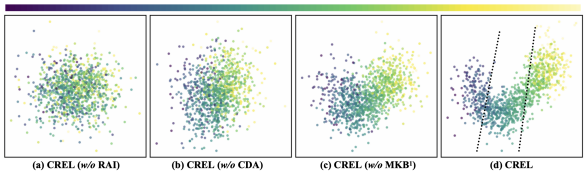

Figure 3: Visualization of feature distribution on MOSI.

Figure 4: Quantitative Analysis on MOSI.

3.6 Visualization of Feature Distribution

To inspect representation geometry under modality incompleteness, we visualize test features on CMU-MOSI using tSNE (Van der Maaten and Hinton, 2008). We extract the fused representation h after retrieval and reliability-aware fusion, as defined in Section 2.3. We use the same tSNE configuration for all settings. Figure 3 compares CREL with representative ablations from Table 4. The variant without retrieval (w/o RAI) shows substantial overlap across sentiment intensities, which indicates weaker separability without external evidence. Removing CDA (w/o CDA) yields a less coherent intensity transition, which is consistent with semantically irrelevant evidence. Replacing prototype memory with instance memory (w/o MKB^1) produces a more dispersed distribution, which suggests noisier evidence units. In contrast, CREL forms a more compact structure with a clearer monotonic

transition along sentiment intensity.

3.7 Quantitative Analysis of Interpretability

We analyze how CREL allocates computation across components during prediction. Let $\alpha_k(x)$ denote the normalized component weight for component k on input x. We obtain $\alpha_k(x)$ from the learned gating modules in CREL. We estimate functional importance using a contribution score based on component removal. For each component k, we compute $I_k(x) = \max\big(0, \mathcal{L}(x; w/o\ k) - \mathcal{L}(x; \text{full})\big)$, where $\mathcal{L}$ is the prediction loss on the same input. We normalize $\{I_k(x)\}$ across components to obtain an importance distribution.

Figure 4 visualizes component weights and importance scores for representative samples. The two distributions show consistent rankings across samples. RAI receives the highest weights and also yields the largest degradation when removed, which matches Table 4. This alignment indicates that the training allocation is consistent with the functional utility of each component. It also supports that each component contributes to robustness under missing inputs.

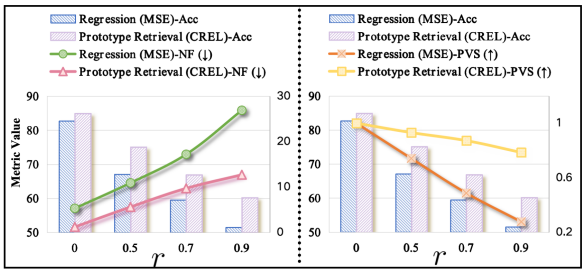

Figure 5: Mean regression artifact diagnostics under intra-modal missingness. Regression collapses toward a near neutral region as r increases, while prototype retrieval preserves dispersion.

3.8 Diagnosing Mean Regression Artifacts

To verify that CREL improves robustness through evidence retrieval rather than regression style completion, we design a targeted diagnostic to quantify mean seeking collapse under missing inputs, thereby supporting CREL's evidence centric advantage under ambiguous supervision. We consider a controlled binary sentiment setting and simulate intra-modal missingness by masking a ratio r of features and replacing masked entries with zeros, with $r \in \{0.0, 0.5, 0.7, 0.9\}$, and defer full details to the appendix. We compare an MSE trained regressor that outputs a continuous prediction score $\hat{s}$ (Regression) with a prototype retrieval baseline that also produces a continuous score $\hat{s}$ for evaluation, where $\hat{s}$ is computed before thresholding and classification is obtained by $\text{sign}(\hat{s})$ (Prototype retrieval). We measure the neutral fraction $\text{NF}_\delta(r) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(|\hat{s}_i^{(r)}| < \delta)$ with $\delta = 0.1$ and the variance shrinkage $\text{PVS}(r) = \frac{\text{Var}(\hat{s}^{(r)})}{\text{Var}(\hat{s}^{(0)})}$. As shown in Figure 5, Regression exhibits pronounced neutral collapse as missingness increases, with $\text{NF}_{0.1}$ rising from 5.3% to 26.9% and PVS dropping from 1.000 to 0.276 at $r = 0.9$, whereas Prototype retrieval mitigates collapse and preserves score dispersion across rates. These observations provide direct evidence that evidence retrieval alleviates mean seeking behavior under missing inputs, consistent with Table 4 where replacing prototype retrieval with regression completion (w/o RAI[2]) yields systematic degradation.

Finally, we provide a Sensitivity Analysis of four main hyperparameters in Appendix H. We also report an Efficiency Analysis against recent state-of-the-art methods in Appendix E. We further report a Stability Analysis on CMU-MOSI (mean±std over five random seeds under missing rates $r \in \{0.0, \ldots, 0.9\}$), see Appendix G.

4 Conclusion

This paper studies multimodal sentiment analysis under modality missingness where incomplete evidence can lead to over-smoothed completion and miscalibrated confidence. We trace these failures to mean-seeking predictions induced by continuous completion under one-to-many affective mappings and to deterministic point estimates under evidence scarcity. We propose CREL, which replaces feature completion with retrieval-based evidence augmentation and evidential inference. CREL retrieves discrete prototype evidence from an inductively constructed Multimodal Knowledge Base, integrates it through retrieval augmented interaction with reliability control, suppresses confounder-related retrieval shortcuts via causal deconfounding adjustment, and calibrates uncertainty using Dirichlet evidential learning. Experiments on three benchmarks under intra-modal and fixed-modality missingness show that CREL improves robustness across diverse settings. Ablations and diagnostics further verify that each component contributes to the overall gains.

Limitations

While CREL improves robustness for multimodal sentiment analysis under uncertain and missing inputs, several limitations remain.

First, our empirical study instantiates CREL in the standard trimodal setting (language, acoustic, and visual). Practical deployments may involve a broader set of modalities, such as infrared, LiDAR, remote sensing, or physiological signals. Extending CREL to such settings will likely require modality-specific backbones and careful adaptation of the representation interfaces to accommodate different spatiotemporal structures and sampling rates. We leave a systematic multi-modality extension and validation for future work.

Second, due to the time and data acquisition costs in this stage, the Multimodal Knowledge Base (MKB) used by CREL is constructed from the training splits of publicly available benchmarks. While this setup is sufficient for controlled evaluation and fair comparison, it limits the coverage and diversity of evidence that the retrieval module can access. In the next stage, we will build a real-world MKB that continuously accumulates multimodal evidence over time, forming a growing and updatable knowledge base rather than a static one.

Third, our experiments focus on sentiment polarity with scalar intensity. Real-world affect also includes fine-grained categorical states (e.g., fear, anger, and joy). Evaluating CREL on fine-grained emotion recognition and multi-label affect annotation is a natural next step.

References

Xun Gao, Xun Jiang, and 1 others. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. 2024. Classifier-guided gradient modulation for enhanced multimodal learning. *Advances in Neural Information Processing Systems*, 37:133328–133344.

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

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*.

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

Yaowei Li, Ruixuan Quan, Lingxi Zhu, and Yi Yang. 2023a. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2604–2613.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023b. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6631–6640.

Ziming Li, Yaxin Liu, Chuanpeng Yang, Yan Zhou, and Songlin Hu. 2025. Rosa: A robust self-adaptive model for multimodal emotion recognition with uncertain missing modalities. *IEEE Transactions on Multimedia*.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432.

Rui Liu, Haolin Zuo, Zheng Lian, Björn W. Schuller, and Haizhou Li. 2024. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 18177–18186.

Chunlei Meng, Jiacheng Yang, Wei Lin, Linqiang Hu, Bowen Liu, Zhuo Zou, LiDa Xu, Zhongxue Gan, and Chun Ouyang. 2025. Multi-grained teacher–student joint representation learning for surface defect classification. *Journal of Industrial Information Integration*, 48:100958.

Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Mohammed E. Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2024a. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325.

Teng Sun, Yinwei Wei, Juntong Ni, Zixin Liu, Xuemeng Song, Yaowei Wang, and Liqiang Nie. 2024b. Multimodal emotion recognition via hierarchical knowledge distillation. *IEEE Transactions on Multimedia*, 26:9036–9046.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, page 6558.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921.

Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. 2024. Gradient-guided modality decoupling for missing-modality robustness. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1083–1092.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, pages 17117–17128.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7617–7630.

Mingzheng Yang, Kai Zhang, Yuyang Ye, Yanghai Zhang, Runlong Yu, and Min Hou. 2025. Decoupling and reconstructing: A multimodal sentiment analysis framework towards robustness. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6803–6811.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10790–10797.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4400–4407.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2023. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, pages 82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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*, pages 2236–2246.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024a. Towards robust multimodal sentiment analysis with incomplete data. *Advances in Neural Information Processing Systems*, 37:55943–55974.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning

language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024b. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618.

Zhao Zhou and 1 others. 2024. Balancing multimodal learning with classifier-guided gradient modulation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22123–22138.

A Related Work

A.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) modeling has been advanced by fusion architectures such as tensor fusion (Zadeh et al., 2017), low rank factorization (Liu et al., 2018), transformer based sequence fusion (Tsai et al., 2019), and pretrained transformer integration (Rahman et al., 2020). Representation factorization has also been explored to separate modality shared and modality specific components for affective learning (Hazarika et al., 2020; Yang et al., 2023). CREL follows this line in targeting complementary evidence, but focuses on missing modality robustness through discrete retrieval and uncertainty calibrated prediction.

A.2 Robust Multimodal Learning

Robustness under missing modalities has been studied in multimodal transformers (Ma et al., 2022). A common line of work performs feature reconstruction and noise aware training (Yuan et al., 2021, 2023) or decouples and reconstructs for robustness (Yang et al., 2025). Proxy driven robustness and cross modal distillation transfer information under incomplete inputs (Zhu et al., 2025; Li et al., 2024; Sun et al., 2024b), and self-adaptive approaches incorporate uncertainty signals during fusion (Li et al., 2025; Gao et al., 2024). Optimization based robustness includes gradient modulation and balancing strategies (Guo et al., 2024; Zhou et al., 2024). In contrast, CREL uses a prototype memory to avoid smooth completion, regularizes retrieval to reduce confounder driven shortcuts, and produces distributional predictions to explicitly represent epistemic uncertainty.

A.3 The Advantages of CREL

At a high level, CREL integrates retrieval, regularization, and evidential modeling. However, its contributions are not a naive stacking of existing components; rather, each component is introduced to address a specific failure mode of missing-modality learning, and they are coupled through a shared retrieval–reliability–evidence pathway.

Retrieval is used to replace smoothing completion, not to augment features. Prior completion-based approaches often rely on continuous regression to reconstruct missing signals, which can encourage mean-regression and over-smoothing under one-to-many supervision. In contrast, CREL retrieves *discrete prototype evidence* from a memory

bank and injects it through reliability-aware interaction, explicitly targeting the collapse phenomenon instead of merely enriching representations.

The deconfounding objective is retrieval-specific and reliability-conditioned. Existing contrastive or regularization methods typically operate on raw features or fused embeddings. CDA is designed around the retrieval process: it constructs counterfactual negatives to suppress shortcut dependencies induced by confounders in the retrieved evidence, and it is explicitly linked to the reliability signal that controls evidence strength. This design aims to reduce spurious retrieval cues that are amplified under missing inputs.

Evidential learning is tied to missingness via controllable evidence strength. Standard evidential learning models uncertainty from logits or learned evidence, but does not explicitly connect uncertainty to the quality of retrieved support under missing modalities. EUC models prediction as Dirichlet evidence and modulates the evidence magnitude by reliability, so that weak support yields higher uncertainty rather than overconfident decisions. This coupling enables a consistent mechanism for uncertainty behavior as missingness increases.

Together, these designs form a coherent solution to missing-modality robustness where retrieval provides candidate evidence, reliability regulates evidence injection, and evidential learning calibrates confidence according to evidence strength.

B Benchmarks

Benchmark	Train	Val	Test	Total
CMU-MOSI	1,284	229	686	2,199
CMU-MOSEI	16,326	1,871	4,659	22,856
CH-SIMS	1,368	456	457	2,281

Table 5: Statistics of Benchmarks.

We used the official training, validation, and test sets for all datasets. Dataset statistics are summarized in Table 5. CMU-MOSI (Zadeh et al., 2016) contains 2,199 utterance-level clips from 93 YouTube videos with sentiment scores in $[-3, 3]$. CMU-MOSEI (Zadeh et al., 2018) contains 22,856 utterance-level clips with sentiment scores in $[-3, 3]$. CH-SIMS (Yu et al., 2020) is a Chinese multimodal sentiment dataset with 2,281 clips from movies and TV series, annotated with sentiment scores in $[-1, 1]$.

C Implementation Details

All experiments are conducted with PyTorch on a single NVIDIA A100 GPU (80GB). We train each model with 5 random seeds and report mean results. We adopt BERT as the language encoder backbone, and we follow the same audio and visual feature extraction pipeline used by the compared baselines to ensure a fair comparison (Li et al., 2025; Zhu et al., 2025). Detailed experimental hyperparameter settings are provided in Table 6.

D Model Zoo

To evaluate CREL under missing-modality settings, we compare it with representative baselines that are actually used in our experiments (Tables 1–3). These baselines span three complementary categories.

General fusion frameworks primarily target standard multimodal fusion and provide strong references for feature extraction and fusion under complete observations. They include DCCAE (Wang et al., 2015), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), CENET (Wang et al., 2022), ALMT (Zhang et al., 2023), and DMD (Li et al., 2023b).

Translation and reconstruction methods explicitly recover missing information via cross-modal translation or feature reconstruction. They include MMIN (Zhao et al., 2021) and CIF-MMIN (Liu et al., 2024) as translation-based approaches, and GCNet (Lian et al., 2023), IMDer (Wang et al., 2023), TFR-Net (Yuan et al., 2021), and DRA (Yang et al., 2025) as reconstruction-oriented solutions.

Robust learning frameworks improve reliability under missingness by uncertainty modeling, noise-robust learning, or knowledge transfer. They include LNLN (Zhang et al., 2024a), NIAT (Yuan et al., 2023), EMT (Sun et al., 2024a), CorrKD (Li et al., 2024), GGMD (Wang et al., 2024), P-RMF (Zhu et al., 2025), and ROSA (Li et al., 2025).

E Efficiency Analysis

Table 7 compares the computational cost of CREL against recent competitive methods on CMU-MOSI. Despite maintaining a parameter scale (117.19M) comparable to lightweight baselines

Hyper-parameter	CMU-MOSI	CMU-MOSEI	CH-SIMS
P, R	256, 8	256, 8	256, 8
$\lambda_{\text{CDA}}, \lambda_{\text{KL}}$	0.1, 0.5	0.1, 0.5	0.1, 0.5
Batch size	128	128	128
Epoch	100	100	100
Optimizer	AdamW	AdamW	AdamW
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Early Stop	✓	✓	✓

Table 6: Hyper-Parameters Setting.

Table 7: Computational efficiency analysis on CMU-MOSI. We report parameter counts and average training time per epoch.

Model	Parameters	Time / Epoch
ConFEDE	246.98M	40.12 s
PRMF	117M	18.62 s
LNLN	116M	24.58 s
CREL (Ours)	117.19M	17.56 s

like PRMF (Zhu et al., 2025) and LNLN (Zhang et al., 2024a), and significantly smaller than ConFEDE (Yang et al., 2023), CREL achieves the lowest per-epoch running time (17.56s). This efficiency stems from its non-generative imputation paradigm: unlike reconstruction-based methods that employ computationally intensive decoders to hallucinate high-dimensional features from scratch, CREL's RAI module performs missing modality completion via efficient vector similarity search over the pre-computed MKB. Since the MKB stores discrete, compact semantic prototypes derived from K-means centroids, the retrieval and fusion process incurs minimal overhead compared to continuous feature generation. Furthermore, the auxiliary modules (CDA and EUC) introduce negligible cost during the forward pass, serving primarily as optimization constraints. Consequently, CREL delivers superior robustness with reduced training latency, confirming that evidential retrieval is a more computationally efficient pathway for modality completion than generative reconstruction.

F Details of Robust Comparison

Following standard protocols for robust MSA, we simulate intra-modal missingness by randomly masking a fraction $r \in \{0.0, 0.1, \ldots, 0.9\}$ of the input sequence during inference. We exclude the trivial case of $r = 1.0$ where no information remains. Detailed performance trajectories for CMU-MOSI, CMU-MOSEI, and CH-SIMS are reported in Tables 8, 9, and 10, respectively. While performance inevitably decays as information sparsity increases, CREL demonstrates a significantly more resilient decay profile compared to generative baselines, particularly in high-missingness regimes ($r \geq 0.7$). This robustness is attributable to the discrete nature of the retrieval-augmented imputation: unlike reconstruction methods that collapse towards the mean when conditioned on sparse inputs, CREL's RAI module retrieves complete, high-fidelity prototypes from the MKB to bridge semantic gaps. Furthermore, the EUC module actively calibrates prediction confidence against evidence sufficiency, effectively suppressing the high-confidence errors that typically inflate MAE in deterministic models under severe data corruption.

G Stability Analysis

We evaluate the training stability of CREL on CMU-MOSI by reporting mean±std under intra-modal missingness. Following the averaged results in Table 1, we select several representative baselines and additionally compute standard deviations for all metrics.

Concretely, for each method we run five trials with different random seeds. In each trial, we evaluate the model under missing rates $r \in \{0.0, 0.1, \ldots, 0.9\}$ and obtain a metric value at each r. We then compute the standard deviation across the five seeds for each r, and finally average these per-r standard deviations over the ten missing-rate settings to obtain a single stability score (std) for each metric. The resulting mean±std statistics are reported in Table 11.

Missing Rate r	CMU-MOSI					
	Acc-2 (%)	F1 (%)	Acc-5 (%)	Acc-7 (%)	MAE (↓)	Corr (%)
0.0	83.25 / 84.82	83.42 / 85.10	50.15	46.25	0.702	0.798
0.1	82.50 / 83.95	82.70 / 84.05	48.92	44.82	0.755	0.762
0.2	80.15 / 81.20	80.25 / 82.45	46.85	42.95	0.798	0.715
0.3	78.45 / 79.10	78.55 / 80.85	45.10	41.50	0.852	0.672
0.4	76.50 / 77.45	76.80 / 79.55	43.25	39.15	0.915	0.635
0.5	73.85 / 74.95	74.15 / 75.35	40.85	37.45	0.985	0.582
0.6	70.25 / 71.15	70.15 / 71.05	36.45	33.25	1.045	0.505
0.7	66.85 / 67.55	66.95 / 66.45	34.25	29.85	1.125	0.415
0.8	62.15 / 63.45	62.15 / 62.05	29.15	24.55	1.255	0.285
0.9	57.75 / 58.98	57.98 / 56.98	23.73	18.13	1.428	0.151
Avg.	**73.17 / 74.26**	**73.31 / 74.39**	**39.87**	**35.79**	**0.986**	**0.552**

Table 8: Robustness evaluation results of CREL on CMU-MOSI under various rates of IMM.

Overall, all compared methods exhibit relatively small standard deviations on most metrics, indicating stable training under random initialization and missingness perturbations. Notably, CREL achieves strong average performance while maintaining competitive (often smaller) variability, suggesting that its gains are not driven by fragile optimization behaviors.

H Sensitivity Analysis

To assess the robustness of CREL to hyperparameter variations, we conduct a sensitivity analysis on CMU-MOSI and CH-SIMS. We vary four key hyperparameters that control the granularity, validity, interaction, and calibration of CREL: (1) the number of prototypes P in MKB, (2) the retrieval width R (Top-R neighbors) in RAI, (3) the de-confounding weight λ_{CDA}, and (4) the KL regularization weight λ_{KL} in evidential learning. We change one parameter at a time and fix the others to $P = 256$, $R = 8$, $\lambda_{\text{CDA}} = 0.1$, and $\lambda_{\text{KL}} = 0.5$.

Impact of Prototype Granularity P. We test $P \in \{64, 128, 256, 512, 1024\}$. Performance improves when increasing P from 64 to 256, as too few prototypes produce over-coarse evidence (underfitting). When P exceeds 512, gains saturate or slightly drop, suggesting that overly fine prototypes may encode instance-specific noise.

Impact of Retrieval Width R. We evaluate $R \in \{1, 2, 4, 8, 16\}$. Aggregating multiple retrieved prototypes ($R = 2$ or 4) consistently outperforms using a single neighbor ($R = 1$), and peak performance is achieved at $R = 8$. Further increasing K (e.g., 16) slightly degrades performance, likely due to long-tail noise from less relevant neighbors. CREL remains stable over $R \in [4, 8]$, showing that reliability estimation and gating effectively suppress noisy evidence.

Impact of De-confounding Weight λ_{CDA}. We sweep $\lambda_{\text{CDA}} \in \{0.01, 0.05, 0.1, 0.2, 0.4\}$ and observe an inverted-U behavior. With too small λ_{CDA} (0.01), retrieval is more affected by confounding shortcuts, reducing evidence quality. The best results occur at $\lambda_{\text{CDA}} = 0.1$. When λ_{CDA} is too large, performance declines due to over-regularization. The overall variance stays within about 1%, indicating limited sensitivity away from extreme values.

Impact of KL Regularization Weight λ_{KL}. We vary the maximum annealing amplitude $\lambda_{\text{KL}} \in \{0.05, 0.1, 0.5, 1.0, 2.0\}$. Too small λ_{KL} provides insufficient regularization and leads to suboptimal calibration under missing evidence, whereas too large λ_{KL} over-regularizes and causes underfitting. We find $\lambda_{\text{KL}} = 0.5$ performs best and slightly exceeds the common choice of 1.0, likely because multimodal fusion benefits from more flexible probability assignment.

Overall, CREL achieves consistently strong performance across reasonable hyperparameter ranges, suggesting its gains mainly come from the architectural design rather than delicate tuning.

Missing Rate r	CMU-MOSEI					
	Acc-2 (%)	F1 (%)	Acc-5 (%)	Acc-7 (%)	MAE (↓)	Corr (%)
0.0	84.15 / 85.85	84.35 / 85.95	53.25	52.15	0.528	0.785
0.1	83.65 / 85.25	83.85 / 85.35	52.85	51.95	0.542	0.768
0.2	83.15 / 84.65	83.45 / 84.75	52.15	51.45	0.558	0.745
0.3	82.25 / 83.55	82.65 / 83.65	51.45	50.95	0.585	0.712
0.4	81.35 / 82.85	81.95 / 82.85	50.85	50.25	0.605	0.685
0.5	80.45 / 81.95	81.25 / 81.95	49.95	49.55	0.635	0.638
0.6	79.55 / 80.55	80.55 / 80.65	48.85	48.55	0.668	0.585
0.7	77.65 / 78.45	78.85 / 78.55	47.15	46.85	0.705	0.505
0.8	73.95 / 72.05	75.35 / 71.75	43.15	44.55	0.765	0.385
0.9	67.65 / 65.55	70.45 / 68.45	39.05	40.35	0.829	0.252
Avg.	**79.38 / 80.07**	**80.27 / 80.39**	**48.87**	**48.66**	**0.642**	**0.606**

Table 9: Robustness evaluation results of CREL on CMU-MOSEI under various rates of IMM.

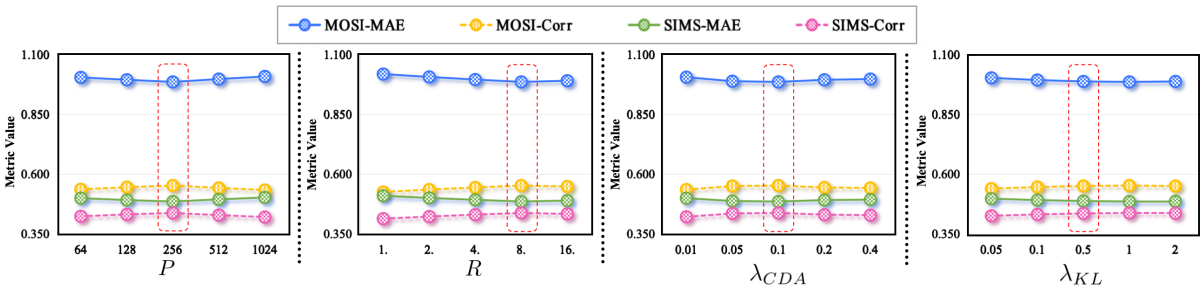

Figure 6: Sensitivity analysis on CMU-MOSI and CH-SIMS under IMM.

I Ethics Statement

This work studies robust multimodal sentiment analysis under missing-modality conditions. All experiments are conducted on publicly available benchmark datasets (e.g., CMU-MOSI, CMU-MOSEI, and CH-SIMS) and we follow the original dataset licenses and usage agreements. Our model is developed for research purposes and aims to improve robustness and uncertainty awareness under incomplete inputs. Potential risks include misuse for profiling or inferring sensitive personal attributes. We do not advocate deploying this model for individual-level decision making, surveillance, or any application that attempts to infer sensitive attributes (e.g., health status, political views, or demographic traits) from personal data. Any downstream use should comply with applicable privacy regulations and obtain informed consent where required. We encourage future work to evaluate bias and fairness under realistic missingness patterns and to incorporate safeguards for responsible deployment.

r	CH-SIMS					
	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
0.0	78.35	81.15	61.25	39.65	0.432	0.565
0.1	77.85	80.95	60.85	39.15	0.438	0.552
0.2	77.25	80.75	60.15	38.65	0.445	0.535
0.3	76.65	80.45	59.45	38.15	0.455	0.515
0.4	76.05	80.15	58.75	37.65	0.465	0.492
0.5	75.15	79.85	58.05	36.85	0.478	0.455
0.6	74.25	79.55	57.15	35.95	0.492	0.415
0.7	73.05	79.25	55.85	34.65	0.515	0.365
0.8	71.25	79.15	54.15	32.55	0.545	0.285
0.9	69.75	79.45	52.25	29.65	0.585	0.191
Avg.	**74.96**	**80.07**	**57.79**	**36.29**	**0.485**	**0.437**

Table 10: Robustness evaluation results of CREL on CH-SIMS under various rates of IMM.

Model	Acc-2	F1	Acc-7	Acc-5	MAE	Corr
MISA	71.49±1.00 / 70.33±0.93	71.28±0.92 / 70.00±1.26	29.85±2.62	33.08±1.77	1.085±0.08	0.524±0.08
Self-MM	70.51±0.71 / 69.26±1.07	66.60±1.92 / 67.54±2.34	29.55±1.06	34.67±1.87	1.070±0.13	0.512±0.07
CENET	71.46±0.60 / 67.73±0.71	68.41±1.12 / 64.85±2.56	30.38±1.27	37.25±1.53	1.080±0.14	0.504±0.11
TFR-Net	68.15±1.25 / 66.35±1.09	61.73±2.82 / 60.06±2.33	29.54±1.00	34.67±1.75	1.200±0.11	0.459±0.25
LNLN	72.55±1.11 / 70.94±1.28	72.73±0.99 / 71.25±1.06	34.26±1.17	38.27±1.23	1.046±0.21	0.527±0.17
CREL	**73.17**±1.02 / **74.26**±1.31	**73.31**±1.07 / **74.39**±1.22	**35.79**±1.06	**39.87**±1.38	**0.986**±0.36	**0.552**±0.13

Table 11: Comparison of model stability on CMU-MOSI. The smaller MAE indicates better performance.