JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

CAMMRF: A Conflict-Aware Missing-Modality Robust Fusion Framework for Multimodal Sentiment Analysis under Heterogeneous Observability

Kajal Chourasiya, Parul Saxena, and Praphula Kumar Jain

Department of Computer Science and Engineering, Madhav Institute of Technology and Science (MITS), Deemed University, Gwalior 474005, Madhya Pradesh, India;
e-mail : kajalchourasiya98@gmail.com; gaurparul2007@mitsgwalior.in; praphulajain@mitsgwalior.in
* Corresponding Author : Kajal Chourasiya

Abstract: Multimodal sentiment analysis combines linguistic, acoustic, and visual evidence, yet deployment commonly violates the benchmark assumption that every modality is present and mutually consistent. Missing-modality systems generally reconstruct absent channels while conflict-aware systems generally assume full observation, even though absence and disagreement both change how far an observed modality should be trusted. We therefore treat the two jointly and hypothesise that conditioning modality reliability on both availability and cross-modal compatibility, and applying it across aligned and disagreement-preserving representations, yields more consistent performance across observability regimes than matched fusion controls without this coupling. We introduce CAMMRF, a framework that couples a mask- and peer-conditioned reliability estimator, an alignment/conflict subspace decomposition, and stochastic modality dropout with cross-view consistency. We evaluate one frozen protocol on the official CMU-MOSI train/validation/test partitions under eight observability regimes; five independent training seeds (42, 43, 44, 45, and 46) completed, each retaining predictions, checkpoints, and logs. In the full-modality regime CAMMRF is best on all four metrics simultaneously—MAE 1.050 ± 0.028, Pearson correlation 0.593 ± 0.009, weighted F1 75.01 ± 1.72%, and Acc-2 74.91 ± 1.80%—rather than on accuracy alone, and it ranks first by mean Acc-2 in 5 of eight regimes. The advantage persists when audio or vision is missing and under the controlled conflict stress (mean Acc-2 73.81%, a 1.10-point decrease, with matching small changes in MAE, weighted F1, and correlation); however, all four metrics degrade together when text is absent—correlation most steeply—so robustness is regime-dependent rather than universal. Full-budget ablations show that several component removals improve full-observation point estimates, supporting an accuracy–robustness trade-off rather than an independent-component claim. The evidence is limited to one corpus and controlled surrogate baselines and does not establish external generalisation. A single Colab notebook regenerates the protocol and exports a machine-readable result registry.

Keywords: Conflict-aware fusion; CMU-MOSI; Missing modality; Multimodal sentiment analysis; Reliability estimation; Robustness; Sentiment analysis; Multimodal fusion.

Received: June, 12th 2026
Revised: August, 12th 2026
Accepted: August, 17th 2026
Published: September, 11th 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licenses (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Multimodal sentiment analysis (MSA) predicts the continuous sentiment expressed in an utterance from language, acoustic delivery, and visual behaviour. It is distinct from text-only sentiment analysis, categorical emotion recognition, and sarcasm detection, although all may share feature extractors. The regression output is commonly accompanied by a binary polarity view obtained from its sign. This paper uses "MSA" only in this sense and reports regression and polarity metrics together.

Progress on CMU-MOSI and CMU-MOSEI has been driven by increasingly expressive fusion mechanisms [1], [2]. Most benchmark experiments nevertheless assume homogeneous observability: text, audio, and vision are all present and their evidence is compatible enough

for a single fusion rule. Deployment violates both halves of this assumption. Privacy filters, recognition failures, and sensor faults remove channels, so the observable set of modalities varies from sample to sample; sarcasm, hesitation, and facial incongruity make the surviving channels disagree, so the trustworthiness of the observable modalities also varies.

These two failure modes are usually studied apart. Missing-modality methods restore or reweight an absent view but implicitly assume that the survivors can be trusted; conflict-aware methods arbitrate disagreement but assume that every channel is present. The two are not independent. Missingness changes which modalities are observable, whereas conflict changes how much the observable ones should be believed. Reconstructing an absent channel from unreliable survivors can amplify their error, and conflict arbitration cannot invoke a channel that has vanished. The research gap therefore concerns a single question that neither line of work addresses on its own: for each sample, how much should an observed modality influence fusion when both the available set and the mutual compatibility of the available modalities change together?

Our rationale for treating the two jointly is that availability and compatibility act on the same latent quantity how far a given modality should be trusted during fusion rather than on unrelated ones. An absent modality is the limiting case of an untrustworthy one: its influence should fall to zero. A present-but-incongruent modality should be down-weighted without being discarded, because its modality-specific signal may still carry the disagreement that matters. A single sample-specific reliability signal, conditioned jointly on the availability mask and on the observed peers, can express this entire continuum from available-and-consistent, through available-but-conflicting, to unavailable. The same signal can then drive both a consensus representation of agreeing evidence and a disagreement-preserving representation of conflicting evidence. This is why we pursue one coupled fusion operator rather than a missing-modality module bolted onto a separate conflict-resolution module.

Accordingly, we hypothesise that a fusion operator in which modality reliability is jointly conditioned on availability and cross-modal compatibility, and is then applied across both aligned and disagreement-preserving representations, will provide more consistent performance across heterogeneous observability regimes than matched fusion controls that do not implement this coupling. We evaluate this hypothesis on CMU-MOSI under heterogeneous observability conditions using complementary quantitative and robustness analyses. The hypothesis is deliberately framed as consistency within the evaluated protocol; it does not assert external or cross-corpus generalisation, which we do not test.

Our objective follows directly. We design and reproducibly evaluate one fusion operator that (i) accepts any non-empty modality subset without reconstructing absent channels, (ii) retains disagreement rather than forcing every representation into one consensus space, and (iii) conditions both behaviours on a common, sample-specific reliability signal. We call the resulting framework Conflict-Aware Missing-Modality Robust Fusion (CAMMRF). The contributions are:

1. a joint problem formulation in which missingness and conflict both change a single peer- and mask-conditioned reliability value, closing the gap left when the two failure modes are handled by separate mechanisms;
2. a dual-stream fusion operator that separates aligned and conflicting subspaces and combines them through the shared reliability signal;
3. a frozen, method-matched evaluation protocol with with component ablations and control checks, with the main evaluation reported on all four metrics;
4. an auditable evidence chain in which all manuscript tables and plots are generated from retained per-seed predictions and a canonical registry; and
5. an explicit boundary between the proposed neural model, deterministic linear controls, and the controlled TFN-S, MulT-S, and MMIM-S surrogates.

The revision makes three evidentiary boundaries explicit. First, the aligned feature bundle verifies source-video disjointness, but it contains no speaker-identity map; speaker-disjointness is therefore not claimed. Second, the interaction baselines are controlled surrogates rather than executions of the original high-capacity architectures. Third, external-corpus generalisation is untested. These boundaries prevent reproducibility from being confused with broader validity.

2. Related Work

2.1. Multimodal sentiment fusion

CMU-MOSI established utterance-level sentiment regression over aligned text, audio, and vision, and CMU-MOSEI enlarged the scale and diversity of the task [1], [2]. Tensor Fusion Network (TFN) explicitly models unimodal, bimodal, and trimodal interactions, while low-rank multimodal fusion reduces the resulting parameter cost [3], [4]. MulT replaces explicit products with directional cross-modal attention for unaligned sequences [5]. MAG-BERT injects non-verbal information into a pretrained language representation [6]; Self-MM generates modality-specific supervision [7]; and MMIM maximises hierarchical mutual information [8]. Contrastive and multi-layer approaches further separate shared and specific information [9]–[11]. These methods motivate interaction modelling but are not designed around arbitrary test-time modality subsets.

2.2. Incomplete multimodal observation

Missing-modality research includes imagination or reconstruction networks, robustness diagnostics, tag-assisted learning, graph completion, dual-level feature restoration, representation transfer, and prompt-conditioned adaptation [12]–[18]. Masked autoencoding and parameter-efficient adaptation extend the same objective to larger backbones [19], [20]. Their central question is how to preserve information when a view is absent. Their reconstruction assumption is useful when observed modalities are trustworthy, but it does not itself decide which survivor should dominate when survivors disagree.

2.3. Conflict and uncertainty

Modality-specific and invariant spaces can preserve information that a single shared embedding removes [21]. Cross-modal graphs and sentiment-aware hierarchical fusion similarly exploit disagreement in sarcasm or related affective tasks [22]–[24]. Recent work studies multi-level conflict, uncertainty-aware fusion, interaction graphs, and foundation-model representations [25]–[30]. The relevant gap is structural: missing-modality methods normally condition on availability, whereas conflict-aware methods normally condition on disagreement. None of the reviewed approaches makes one reliability value depend jointly on the availability mask and the observed peers and then uses that value in both consensus and disagreement streams.

2.4. Positioning and novelty boundary

CAMMRF adopts modality encoders, masking, attention, and consistency regularisation as established ingredients. It adapts shared/private decomposition to an alignment/conflict interpretation. The claimed novelty is their coupling: the same mask- and peer-conditioned reliability score gates the aligned stream, reads the conflict stream, and responds to the masks used by cross-view training. We do not claim invention of the individual ingredients, superiority to published leaderboard systems, or external generality. The controlled study isolates this coupling under matched pooled features and evaluation regimes.

2.5. Relationship to prior work and research gap

Three method families frame CAMMRF conceptually rather than through the experimental values reported later. Interaction models such as Tensor Fusion, MulT, and MMIM were designed around sequence-level neural objectives and larger contextual representations [3], [5], [8]; they motivate expressive fusion but are not built for arbitrary test-time modality subsets. In this controlled study their interaction mechanisms are represented by budget-matched surrogates (TFN-S, MulT-S, MMIM-S) so that a missingness comparison remains internally traceable. This choice deliberately narrows any architectural claim: it cannot establish that CAMMRF would outperform the official high-capacity implementations, and the comparison is confined to the shared pooled-feature and ridge-read-out budget.

The missing-modality family typically restores or imagines an absent representation [12], [14]–[16]. CAMMRF instead removes the absent view and reweights the survivors, which avoids reconstruction error and, by construction, lets one model serve every non-empty modality subset. A structural expectation follows conceptually: when the most informative

channel is itself absent, no reconstruction or reweighting can manufacture the missing semantic content, so reliability weighting should help most when an informative channel survives and least when it does not. Prior work conditions such repair on availability alone.

The conflict-aware family preserves modality-specific information so that incongruity is not averaged away [21], [25]. CAMMRF shares this goal through a disagreement-preserving stream, but conditions it on the same availability- and peer-derived reliability signal that governs the consensus stream. Prior work conditions such arbitration on disagreement alone. The distinguishing property relative to both families is therefore the coupling: none of the reviewed approaches makes one reliability value depend jointly on the availability mask and the observed peers and then uses that value in both the consensus and the disagreement representation.

This positioning closes the research gap stated in Section 1: availability and compatibility should modulate a single sample-specific trust signal instead of two disconnected mechanisms. The next section formalises that signal and the dual-stream operator that consumes it; whether the resulting behaviour matches these conceptual expectations is an empirical question that Section 5 addresses on all four evaluation metrics.

3. Methodology

3.1. Problem definition

For utterance n, let the aligned modalities be $X_i \in \mathbb{R}^{(L\times d_i)}$ for $i \in M = \{T, A, V\}$, with $L = 50$ and dimensions $(d_T, d_A, d_V) = (300,5,20)$. A binary mask $m_i \in \{0,1\}$ states whether modality $i$is observed, with $\sum_i m_i \geq 1$. We define the combined mask vector $m = (m_T, m_A, m_V) \in 0,1^3$. The target $y \in [-3,3]$ is a continuous sentiment intensity. The model estimates $\hat{y} = f(X_T, X_A, X_V, \mathbf{m})$ without imputing a missing input. Missing inputs are zeroed before and after biased projections, and attention excludes absent keys and values.

3.2. Forward pass and component configuration

The implementation performs the following five steps, with the implemented forward-pass configuration summarized in Table 1 and illustrated in Figure 1.

1. Pool and encode: Each aligned sequence is mean-pooled, standardised with training-only statistics, and mapped by a two-layer modality encoder:

$$\mathbf{h}_i = \mathrm{Enc}_i(\mathrm{Std}(1/L \sum_{\ell=1}^{L} X_{i\ell})), \qquad \mathbf{h}_i \in \mathbb{R}^{64}. \tag{1}$$

2. Estimate reliability: For each observed modality, the peer context is the mean representation of the other observed modalities. A small network receives $\mathbf{h}_i$, that peer context, and m: Here, the peer context is denoted by $\bar{h}_{-i}$. A missing channel is assigned reliability zero; an observed channel may be downweighted when it disagrees with its observed peers.

$$r_i = m_i\, \sigma\big(g_i([\mathbf{h}_i; \bar{\mathbf{h}}_{-i}; \mathbf{m}])\big). \tag{2}$$

3. Separate aligned and conflicting information: Learned maps produce an alignment component $\mathbf{a}_i = A_i \mathbf{h}_i$ (2a) and a conflict component $\mathbf{c}_i = C_i \mathbf{h}_i$ (2b). An observed-only orthogonality term discourages the two components from collapsing to the same representation.
4. Fuse two streams: The alignment stream is a reliability-weighted consensus,

$$\mathbf{z}_a = \frac{\sum_i r_i\, \mathbf{a}_i}{\epsilon + \sum_i r_i}. \tag{3}$$

The conflict stream applies masked attention to the $\mathbf{c}_i$ values and uses reliability in readout. Their concatenation is mapped to y.

5. Train across views A non-empty random mask is sampled during training. The masked and full views share parameters; a cosine–MSE consistency term connects their fused representations.

Table 1. Implemented forward-pass configuration. All learned blocks belong to the single CAMMRF model and are trained jointly.

Block	Inputs	Implemented role
Unimodal encoders	Pooled $T/A/V$ features	Two-layer projections to 64-dimensional representations, with masking before and after biased maps
Reliability estimator	$\mathbf{h}_i$, observed-peer mean, m	Produces $r_i \in [0,1]$ and forces $r_i = 0$ for an absent modality
Alignment maps	Observed $\mathbf{h}_i$	Construct the reliability-weighted consensus stream
Conflict maps and attention	Observed $\mathbf{h}_i$, attention mask	Preserve and read out disagreement without attending to absent keys or values
Fusion/read-out	Alignment and conflict streams	Concatenate both streams and predict continuous sentiment
Consistency branch	Full and stochastic masked views	Couple their fused representations during training only

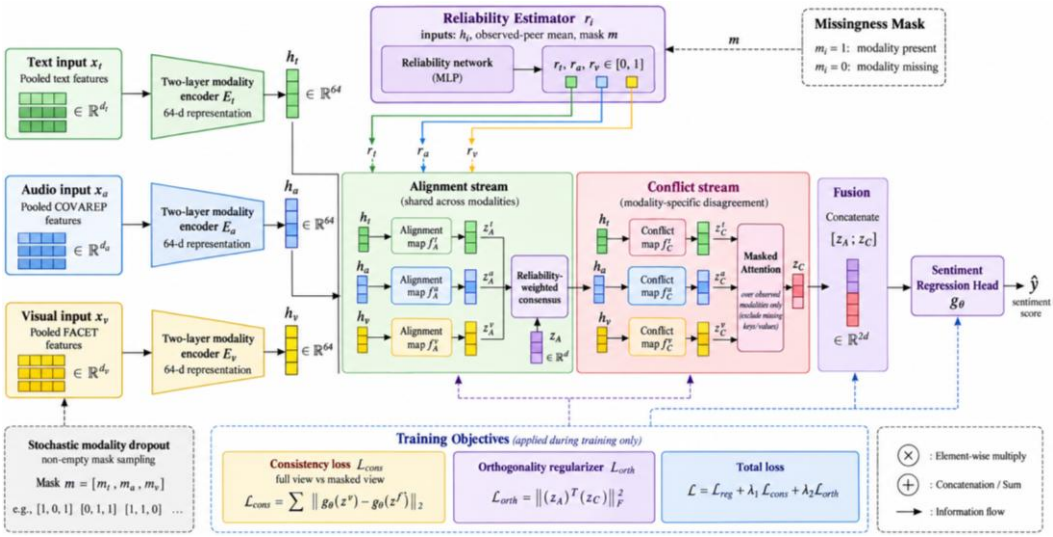


Figure 1. Architecture of CAMMRF. The availability mask and observed peer context jointly determine reliability; aligned and conflict streams remain separate until the final fusion.

3.3. Training objective

The task loss is mean squared error on the normalised training target. Regularisers address cross-view consistency, within-view alignment, conflict preservation, and observed-only orthogonality:

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \beta_{\text{cons}}\mathcal{L}_{\text{cons}} + \beta_{\text{align}}\mathcal{L}_{\text{align}} + \beta_{\text{conf}}\mathcal{L}_{\text{conf}} + \beta_{\text{orth}}\mathcal{L}_{\text{orth}}. \quad (4)$$

The frozen coefficients are 0.30, 0.10, 0.05, and 0.02, respectively. An auxiliary MSE coefficient of 0.10 is retained in the implemented consistency branch. Stochastic masking is applied with probability 0.50 and samples uniformly from the seven non-empty three-modality masks. These values were frozen before the five-seed run and are recorded in the configuration hash.

3.4. Implementation boundary

The evaluated CAMMRF instance has 96,260 trainable parameters. This number is read from the executed run manifest, not inferred from the diagram. The architecture accepts a known mask at inference. It does not contain a missingness detector, temporal sequence encoder, reconstruction network, or modality-specific fallback head. Those omissions matter in the single-channel regimes and are retained as limitations rather than obscured by additional components.

3.5. Why the mechanisms are coupled

Reliability is not an auxiliary score reported after fusion. It changes the representation that reaches both branches. If a modality is absent, its mask removes it from the peer mean, sets its reliability to zero, and excludes it from attention. If it is present but atypical relative to its observed peers, the estimator can reduce its aligned contribution while the private conflict representation remains available. The stochastic training masks expose the same estimator to changing peer sets, and the consistency term discourages the fused representation from changing arbitrarily when a view is removed. The intended behaviour is therefore a continuum from available-and-consistent to available-but-conflicting to unavailable, rather than a sequence of independent repair modules.

This coupling also defines the ablation interpretation. Disabling reliability changes both streams; disabling alignment or conflict changes which representation can use the reliability value; disabling dropout or consistency changes which observability patterns shape that value during training. The ablations are consequently tests of interacting mechanisms, not orthogonal estimates of additive component effects.

4. Experimental Setup

4.1. Dataset, partitions, and features

The study uses one aligned CMU-MOSI feature bundle. Its verified SHA-256 is shown below so that the exact input can be checked independently: 1a113ad5edc8b9b625a7e29e6b94a97ecade70494c4e022d9f7ba5affca3bbdc

The train, validation, and test partitions contain 1,284, 229, and 686 utterances. Their source-video counts are 52, 10, and 31, with no source-video overlap across partitions. The bundle stores segment identifiers but no speaker-identity map; consequently, source-video disjointness is verified and speaker-disjointness remains unverified. This replaces the stronger unsupported claim that each video identifier necessarily denotes a distinct speaker.

Text features have width 300, acoustic COVAREP features width 5, and visual FACET features width 20 over 50 aligned steps [31]–[33]. Each sequence is mean-pooled. Per-feature mean and population standard deviation are estimated on the training partition only and applied unchanged to validation and test; target normalisation also uses training statistics. All tensors are checked for shape and finite values. The test labels contain 30 exact neutrals, so MAE and correlation use all 686 test utterances while Acc-2 and weighted F1 use 656 non-neutral utterances.

Table 2. Validated CMU-MOSI feature-bundle composition. Source-video counts come from the first column of the stored $N, 3$ identifier array

Partition	Utterances	Source videos	Neutral	Non-neutral
Train	1,284	52	53	1,231
Validation	229	10	13	216
Test	686	31	30	656

4.2. Observability regimes

All methods are evaluated on the same eight test conditions: Full; text, audio, or vision missing; text-, audio-, or vision-only; and Controlled Perturbation. Missing modalities are zeroed under the same mask for every method. The Controlled Perturbation condition adds Gaussian noise with standard deviation 0.35 to the standardised full-modality features. One perturbation realisation, generated with seed 44, is fixed across methods and training seeds. This controls the stress input but should not be interpreted as a naturalistic cross-modal conflict condition.

4.3. Baselines and fairness boundary

Early Fusion concatenates the three pooled views and fits a validation-selected ridge read-out. Late Fusion fits three unimodal ridge models and averages the observed predictions. Both are deterministic for the fixed data and therefore have zero across-seed variation.

TFN-S, MulT-S, and MMIM-S are controlled surrogates. TFN-S applies seeded modality projections followed by Hadamard interaction features; MulT-S uses seeded directional interaction features; MMIM-S uses seeded shared and modality-specific projected features. Each has the same training-only standardisation, validation-selected ridge grid {0.1,1,10,100}, masks, conflict input, split, and metrics. Stochastic projections are regenerated for every training seed. The suffix "-S" is retained because these are compact feature-map controls, not the published TFN, MulT, or MMIM architectures. Their values must not be compared to published leaderboard results as if implementation fidelity were equivalent.

CAMMRF receives the availability mask internally. A separate fairness test refits the controlled surrogates and the deterministic controls with and without explicit mask indicators. The average effect is small and slightly negative, so mask access alone does not explain the observed ranking.

Table 3. Baseline-fidelity and mask-access boundary. "Seeded" means the feature map is regenerated for each training seed.

Method	Status	Stochastic element	Fairness boundary
Early Fusion	Deterministic control	None	Same pooled features, ridge grid, validation criterion, masks, and test regimes
Late Fusion	Deterministic control	None	Unimodal ridge models averaged over observed channels
TFN-S	Controlled surrogate	Seeded projections	Hadamard interaction map; not the published neural TFN
MulT-S	Controlled surrogate	Seeded projections	Directional interaction map; not the published transformer
MMIM-S	Controlled surrogate	Seeded projections	Shared/private feature map; not the published MMIM objective
CAMMRF	Proposed neural model	Initialisation, batches, masks	Uses the same data and regimes but a validation-selected neural checkpoint

4.4. Optimisation and reproducibility

CAMMRF is trained from scratch for seeds 42, 43, 44, 45, and 46 with AdamW, learning rate 3×10^{-3}, weight decay 10^{-4}, batch size 64, maximum 40 epochs, five warm-up epochs, cosine decay, gradient clipping at 1.0, input dropout 0.20, and early-stopping patience 8. The checkpoint with the lowest validation MAE is restored. Python, NumPy, and PyTorch are seeded; deterministic operations are requested where supported. All five runs completed on a Tesla T4 under Python 3.12.13, PyTorch 2.11.0, CUDA 12.8. Each run retains the checkpoint, training history, projection hashes, predictions for six methods and eight regimes, metrics, and a success manifest. The exact dataset and configuration hashes are checked before aggregation.

Table 4. Frozen training and evaluation protocol

Setting	Value	Setting	Value
Seeds	42–46	Hidden width	64
Optimizer	AdamW	Learning rate	3×10^{-3}
Weight decay	10^{-4}	Batch size	64
Maximum epochs	40	Early-stop patience	8
Warm-up epochs	5	Input dropout	0.20
Gradient clip	1.0	Mask sampling	non-empty, α=0.50
Selection metric	validation MAE	Conflict σ/ seed	0.35 / 44

4.5. Metrics and statistical analysis

MAE and Pearson correlation evaluate continuous sentiment; lower MAE and higher correlation are better. Acc-2 maps each non-neutral target and prediction to its sign. Weighted

F1 uses the same non-neutral subset and weights class F1 values by class support. This pairing guards against a polarity score that is inflated by class imbalance.

Across-seed results are reported as mean ± sample standard deviation; per-seed values remain available. For the original seed, paired absolute-error differences are averaged within the 31 test source videos before the paired t-test and Wilcoxon signed-rank test. Confidence in-tervals resample source-video clusters with replacement. Cohen's d uses the cluster-mean differences. Friedman tests rank the six methods across the eight regimes, followed by a Nemenyi post-hoc analysis using a critical difference. Because these regimes are related per-turbations of one corpus rather than independent datasets, the rank analysis is reported as descriptive support rather than as evidence of cross-dataset generality.

4.6. Execution and evidence retention

The canonical notebook runs environment checks, dataset validation, preprocessing, self-tests, training, baseline fitting, eight-regime prediction, ablation, statistics, figure and table generation, aggregation, manifest creation, and archive packaging in order. A seed is included in the aggregate only after its success manifest and required artifacts exist. The final evidence contains 6×8×5=240 prediction archives. Checkpoint, projection, dataset, and configuration hashes connect each aggregate value to a primary run. The canonical CSV contains 192 method–regime–metric aggregates and the per-seed registry contains 960 rows. No manuscript result table is maintained as an independent handwritten copy.

5. Results and Discussion

5.1. Five-seed full-modality results

Table 5 summarizes the full-modality performance across five completed seeds and four complementary evaluation metrics. CAMMRF achieves the best result on all four metrics, with an MAE of 1.050 and a Pearson correlation of 0.593 on the continuous sentiment target. These correspond to improvements of 0.167 in MAE and 0.081 in correlation over Early Fusion. For polarity classification, CAMMRF also obtains the highest Acc-2 (74.91%) and weighted F1 (75.01%), exceeding Early Fusion by 2.04 and 2.17 percentage points, respectively. The consistent ranking across regression and classification measures indicates that the observed advantage is not confined to a single evaluation criterion

Table 5. Full-modality test performance over five seeds (mean ± sample SD). Acc-2 and weighted F1 use 656 non-neutral utterances; MAE and Corr use all 686.

Method	MAE ↓	Acc-2 (%) ↑	F1 (%) ↑	Corr ↑
CAMMRF	1.050 ± 0.028	74.91 ± 1.80	75.01 ± 1.72	0.593 ± 0.009
Early Fusion	1.217 ± 0.000	72.87 ± 0.00	72.84 ± 0.00	0.512 ± 0.000
Late Fusion	1.265 ± 0.000	61.89 ± 0.00	60.28 ± 0.00	0.522 ± 0.000
TFN-S	1.336 ± 0.036	59.48 ± 2.15	58.75 ± 2.34	0.352 ± 0.056
MulT-S	1.316 ± 0.030	61.16 ± 2.52	60.31 ± 2.96	0.377 ± 0.041
MMIM-S	1.265 ± 0.034	66.68 ± 1.65	66.38 ± 1.73	0.465 ± 0.040

The variability across seeds is relatively small for CAMMRF, particularly for Pearson correlation (SD = 0.009), although larger variation is observed for Acc-2 and weighted F1. In contrast, Early Fusion and Late Fusion have zero across-seed SD because their closed-form solutions do not involve stochastic projection. These zero values therefore reflect the deterministic nature of the corresponding controls rather than an absence of uncertainty in their performance. Overall, the full-modality results establish the reference performance for evaluating the broader hypothesis under heterogeneous observability. The subsequent analyses therefore examine whether the observed performance is retained when modality availability changes or when the input is subjected to controlled cross-modal perturbation.

5.2. Heterogeneous observability

Table 6 reports CAMMRF performance across the eight observability regimes using the same four metrics as the full-modality evaluation. The results show a clear dependence on the

availability of text. When either audio or vision is removed, performance remains close to the Full condition across all metrics. Audio missing changes Acc-2, MAE, weighted F1, and correlation only marginally, and the visual-missing condition shows a similar pattern. These results indicate that, within this pooled-feature setting, the remaining modalities can largely preserve the performance obtained under complete observation.

Table 6. CAMMRF performance in all eight regimes over five seeds (mean ± sample SD).

Regime	Acc-2 (%)	MAE	F1 (%)	Corr
Full	74.91 ± 1.80	1.050 ± 0.028	75.01 ± 1.72	0.593 ± 0.009
Text missing	52.32 ± 2.35	1.435 ± 0.033	49.09 ± 4.91	0.185 ± 0.022
Audio missing	74.85 ± 1.68	1.066 ± 0.032	74.95 ± 1.63	0.586 ± 0.012
Visual missing	75.09 ± 1.25	1.066 ± 0.025	75.17 ± 1.14	0.591 ± 0.008
Text only	74.48 ± 1.21	1.130 ± 0.057	74.60 ± 1.17	0.571 ± 0.012
Audio only	53.38 ± 2.47	1.406 ± 0.027	50.70 ± 4.53	0.196 ± 0.039
Visual only	50.85 ± 2.57	1.476 ± 0.063	47.23 ± 5.68	0.143 ± 0.029
Conflict stress	73.81 ± 1.05	1.084 ± 0.026	73.91 ± 0.97	0.561 ± 0.009

The effect is substantially different when text is unavailable. As shown in Table 6, Text missing reduces Acc-2 to 52.32%, weighted F1 to 49.09%, and correlation to 0.185, while MAE increases to 1.435. The Audio only and Visual only regimes exhibit a comparable deterioration. In contrast, Text only retains relatively strong polarity performance, with Acc-2 and weighted F1 remaining close to the Full condition, although MAE and correlation still deteriorate. This difference suggests that the text modality carries a particularly important role in the pooled representation: it can support polarity discrimination when used alone, whereas the absence of text substantially limits the model's ability to recover continuous sentiment information from the remaining channels.

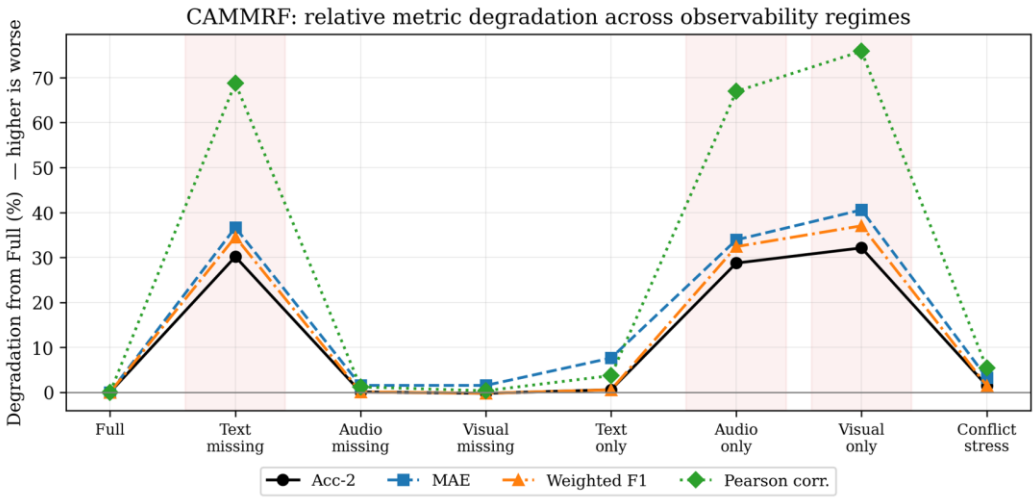


Figure 2. Relative degradation of each metric from the full-modality reference across the eight regimes (five-seed means). Higher values indicate greater degradation; MAE is represented by its increase, while Acc-2, weighted F1, and Pearson correlation are represented by their decrease.

The relative changes are more apparent in Figure 2. The figure shows that the text-bearing regimes remain close to the Full reference, whereas the text-absent regimes produce substantially larger degradation across all four metrics. Pearson correlation shows the steepest relative decline, highlighting a loss of agreement with the continuous sentiment target that is less apparent from Acc-2 alone. Thus, the multi-metric view provides a more complete characterization of the effect of modality removal than polarity accuracy alone. Under Conflict stress, the performance remains relatively close to the Full condition: Acc-2 and weighted F1 decrease to 73.81% and 73.91%, respectively, while MAE increases to 1.084 and correlation decreases to 0.561 (Table 6). As illustrated in Figure 2, the degradation is small compared with

the text-absent regimes. However, this result should be interpreted narrowly because the conflict condition represents a single controlled synthetic perturbation rather than the broader range of cross-modal disagreement that may occur in naturally collected data.

5.3. Run-to-run variation

Table 7 reports the full-modality results for each of the five completed training seeds. The variation across runs is relatively modest, with MAE ranging from 1.005 to 1.075 and Pearson correlation from 0.582 to 0.604. Acc-2 and weighted F1 show somewhat larger variation, ranging from 73.63% to 77.90% and from 73.79% to 77.86%, respectively. The strongest classification results occur for seed 43, whereas seed 46 produces the highest correlation, indicating that the relative ordering of individual runs can differ slightly across evaluation criteria.

Table 7. Retained CAMMRF full-modality results for every completed training seed, reported for all four metrics.

Seed	MAE	Acc-2 (%)	F1 (%)	Corr
42	1.058	73.93	74.06	0.582
43	1.005	77.90	77.86	0.598
44	1.075	73.78	73.91	0.596
45	1.068	73.63	73.79	0.587
46	1.044	75.30	75.45	0.604

Taken together, the results in Table 7 are consistent with the aggregate statistics in Table 5 and do not indicate a single anomalous training run driving the full-modality result. However, the five-seed sample remains limited, so the reported standard deviations should be interpreted as estimates of run-to-run variability within this experimental protocol rather than as population-level uncertainty.

5.4. Full-budget ablation

Table 8 examines the effect of removing individual components from the complete CAMMRF configuration under a common training budget. The results do not indicate that every component independently improves full-observation performance. In particular, removing consistency produces the highest Full Acc-2 (76.07%) and the lowest Full MAE (1.047), while the other ablated variants also achieve Full Acc-2 values at or above the reference configuration. This suggests that the contribution of individual components cannot be evaluated solely from complete-modality performance.

Table 8. Seed-42 ablation with the same 40-epoch maximum and patience 8 for every variant. One-missing Acc-2 averages text, audio, and vision missing.

Variant	Full Acc-2 (%)	Full MAE	One-missing Acc-2 (%)
Ablation reference (full-budget)	73.93	1.058	67.53
w/o Reliability (R)	75.30	1.070	68.60
w/o Alignment (A)	74.70	1.081	68.34
w/o Conflict (C)	75.15	1.084	65.96
w/o Consistency (S)	76.07	1.047	67.33
w/o Dropout (D)	74.70	1.059	65.29

The pattern changes under incomplete observation. As shown in Table 8, removing conflict reduces the one-missing Acc-2 to 65.96%, while removing modality dropout produces a larger reduction to 65.29%. In contrast, removing reliability or alignment slightly increases the one-missing point estimate relative to the reference. The contrast between the full and one-missing results indicates that some components may contribute more to robustness under incomplete observation than to peak performance when all modalities are available.

These results should be interpreted as evidence of component interaction rather than isolated causal effects, because the ablations are jointly retrained variants evaluated on a single

seed and the available ablation record covers only Full Acc-2, Full MAE, and one-missing Acc-2. A complete four-metric ablation analysis is therefore not available. Within this scope, the deployed configuration represents a compromise between full-observation performance and robustness to missing modalities rather than a configuration in which every individual component is uniformly superior.

5.5. Paired inference and rank analysis

Table 9 reports the cluster-aware paired comparison between CAMMRF and the two fusion controls using absolute error on the retained seed-42 predictions. CAMMRF achieves lower error than both Early Fusion and Late Fusion. Relative to Early Fusion, the mean difference in absolute error is −0.159, with a 95% cluster confidence interval of [−0.231, −0.086], Wilcoxon p = 0.0010, and Cohen's d = −0.65. Against Late Fusion, the corresponding difference is −0.207, with a 95% cluster confidence interval of [−0.311, −0.109], p = 0.0004, and Cohen's d = −0.72. Thus, both comparisons show confidence intervals that exclude zero and moderate effect sizes, providing inferential support for the lower prediction error observed for CAMMRF in the full-modality setting.

Table 9. Seed-42 paired absolute-error comparisons. Negative ΔMAE favours CAMMRF. Tests and intervals use 31 source-video clusters; point estimates retain utterance weighting.

Comparison	ΔMAE	95% cluster CI	p_t	Cohen's d
CAMMRF vs Early Fusion	-0.159	[-0.231, -0.086]	0.0010	-0.65
CAMMRF vs Late Fusion	-0.207	[-0.311, -0.109]	0.0004	-0.72

The analysis is intentionally restricted to absolute error because this metric permits the cluster-aware paired comparison used here, while the other evaluation metrics are retained for the descriptive multi-metric analysis. The inference therefore complements rather than replaces the five-seed results in Table 5. In particular, the statistical test is based on the retained seed-42 predictions and 31 source-video clusters, whereas Table 5 summarizes performance across five training seeds. The two analyses consequently address different aspects of the evidence: the former evaluates whether the observed paired error difference is supported after accounting for source-video clustering, while the latter characterizes run-to-run performance across training seeds.

Rank analysis provides a broader comparison across the heterogeneous observability protocol. Across the eight seed-42 regimes, Friedman tests reject equal ranks for MAE, Pearson correlation, Acc-2, and weighted F1, with the Nemenyi analysis using the reported critical difference for subsequent pairwise rank interpretation. This result is consistent with the broader pattern observed across the four metrics rather than being driven by a single evaluation criterion. However, the eight regimes are related perturbations of the same test set rather than independent datasets. The resulting rank statistics should therefore be interpreted as evidence of consistent within-protocol ordering, not as evidence of external generalisation or independent-dataset superiority.

5.6. Mask access, interpretation, and practical boundary

The mask-access control examines whether the observed missing-modality performance could be explained simply by providing the competing methods with explicit availability information. When the availability mask is exposed to the controlled surrogates and deterministic controls, the resulting changes in missing-modality Acc-2 are small: Early Fusion, TFN-S, MMIM-S, and MulT-S change by the reported amounts, with an average change of −0.13 percentage points across the four controls. This indicates that access to the mask alone does not reproduce the observed comparison. The result is consistent with the possibility that the benefit comes from how CAMMRF incorporates availability information into reliability and fusion, although the control does not isolate a unique mechanism.

The practical interpretation is therefore narrower than a general claim of robustness. CAMMRF uses the same reliability mechanism across the evaluated non-empty modality configurations and remains relatively stable under the controlled conflict stress, but its performance is strongly dependent on which information remains available. In particular, the text-missing and non-text single-channel conditions represent clear boundaries of the current

approach. In a deployed setting, additional mechanisms would be needed to handle errors in missingness detection and calibration, while temporal encoders and evaluation on more diverse multimodal corpora would be necessary to assess whether the observed behaviour extends beyond the current pooled-feature protocol.

5.7. Integrated synthesis of findings

Taken together, the experiments provide partial support for the hypothesis formulated in Section 1. The full-modality results show that CAMMRF performs favourably across all four metrics, while the heterogeneous-observability analysis shows that this advantage is largely retained when audio or vision is unavailable and under the controlled conflict condition. The evidence is substantially weaker when text is absent, indicating that reliability-aware fusion can regulate the contribution of available modalities but cannot fully compensate for the loss of a modality carrying substantial semantic information.

The different analyses also clarify what can and cannot be attributed to the proposed design. The paired analysis in Table 9 supports lower seed-42 absolute error relative to both Early Fusion and Late Fusion, while the rank analysis indicates consistent within-protocol ordering across the four metrics. The mask-access control further shows that simply exposing availability information to the competing methods does not reproduce the observed missing-modality comparison. These results are consistent with the intended role of the coupled reliability mechanism, but they do not by themselves establish that any individual component is solely responsible for the overall performance.

This interpretation is particularly important in light of the ablation results. Several component removals improve full-observation point estimates, whereas removing the conflict branch or modality dropout reduces one-missing performance. The results therefore point to an accuracy–robustness trade-off rather than a set of components that each provide an independent improvement. The retained CAMMRF configuration should consequently be understood as a compromise across observability conditions rather than as uniformly optimal under every individual metric or regime.

Finally, the strength of the conclusions is bounded by the evaluation protocol. The evidence comes from a single CMU-MOSI feature bundle and related perturbations of one test set; several interaction baselines are controlled surrogate implementations, the availability mask is assumed to be known, and the ablation analysis is based on seed 42 with only a subset of metrics retained. The five-seed results provide evidence of run-to-run consistency but remain limited in scope. Accordingly, the experiments establish internal, multi-metric support for the proposed reliability-aware fusion strategy within the evaluated protocol, but they do not establish external cross-corpus generalisation, superiority over faithful implementations of the original high-capacity baselines, robustness to imperfect modality detection, or effectiveness when the dominant text modality is unavailable.

6. Conclusions

This study investigated whether missing modalities and cross-modal disagreement can be jointly incorporated into sample-specific reliability estimation for multimodal sentiment analysis. CAMMRF addresses this problem by coupling availability- and peer-conditioned reliability with aligned consensus and disagreement-preserving representations within a unified fusion framework. Under the evaluated CMU-MOSI protocol, the findings provide partial support for this approach: the framework performs consistently across complete and several incomplete or conflicting observation settings, while its effectiveness remains dependent on the information retained by the available modalities.

The contribution of this work is primarily methodological. It shows how modality availability and cross-modal compatibility can be incorporated into a common reliability-aware fusion mechanism without introducing a separate modality-imputation network. The results further suggest that robustness to heterogeneous observations should be considered alongside complete-modality performance when evaluating multimodal fusion, since a configuration that performs well under full observation need not provide the same balance under incomplete observations.

The conclusions are bounded by the scope of the current evaluation. The experiments use a single CMU-MOSI feature bundle with mean-pooled modality representations and a known availability mask, while TFN-S, MulT-S, and MMIM-S serve as controlled surrogate

baselines rather than faithful implementations of their original architectures. The stochastic and ablation analyses are also limited in scope. Consequently, the present study does not establish speaker-disjoint or cross-corpus generalisation, robustness to imperfect modality detection, or superiority over faithful implementations of the original baselines. Future work should extend the evaluation to additional multimodal corpora and temporal representations, consider realistic and uncertain missingness, and strengthen baseline, ablation, and multi-seed validation to assess the broader applicability of the proposed framework.

Author Contributions: Conceptualisation: K. Chourasiya; Methodology: K. Chourasiya; P. K. Jain; Software: K. Chourasiya; Investigation: K. Chourasiya; Visualisation: K. Chourasiya; Original draft: K. Chourasiya; Validation: P. Saxena; P. K. Jain; Supervision: P. Saxena; P. K. Jain; Review: P. Saxena; P. K. Jain. All authors are responsible for approving the final submitted version.

Funding: The authors received no specific funding for this research from a public, commercial, or not-for-profit agency.

Data Availability Statement: : This study uses the publicly distributed CMU-MOSI benchmark [1]. No new data were collected.

Acknowledgments: The authors acknowledge the use of AI-assisted tools for language editing and manuscript proofreading. All research design, analysis, interpretation, and final content were independently reviewed and validated by the authors.

Conflicts of Interest: The authors declare no known competing financial interests or personal relationships that could have influenced this work.

References

[1] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv Prepr. arXiv1606.06259*, 2016.

[2] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "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)*, 2018, pp. 2236–2246. doi: 10.18653/v1/P18-1208.

[3] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor Fusion Network for Multimodal Sentiment Analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114. doi: 10.18653/v1/D17-1115.

[4] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient Low-rank Multimodal Fusion With Modality-Specific Factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2247–2256. doi: 10.18653/v1/P18-1209.

[5] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569. doi: 10.18653/v1/P19-1656.

[6] W. Rahman *et al.*, "Integrating Multimodal Information in Large Pretrained Transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2359–2369. doi: 10.18653/v1/2020.acl-main.214.

[7] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, pp. 10790–10797, May 2021, doi: 10.1609/aaai.v35i12.17289.

[8] W. Han, H. Chen, and S. Poria, "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*, 2021, pp. 9180–9192. doi: 10.18653/v1/2021.emnlp-main.723.

[9] S. Yang, L. Cui, L. Wang, and T. Wang, "Cross-modal contrastive learning for multimodal sentiment recognition," *Appl. Intell.*, vol. 54, no. 5, pp. 4260–4276, Mar. 2024, doi: 10.1007/s10489-024-05355-8.

[10] Z. Li, B. Xu, C. Zhu, and T. Zhao, "CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 2282–2294. doi: 10.18653/v1/2022.findings-naacl.175.

[11] Y. Cai, X. Li, Y. Zhang, J. Li, F. Zhu, and L. Rao, "Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning," *Sci. Rep.*, vol. 15, no. 1, p. 2126, Jan. 2025, doi: 10.1038/s41598-025-85859-6.

[12] J. Zhao, R. Li, and Q. Jin, "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)*, 2021, pp. 2608–2618. doi: 10.18653/v1/2021.acl-long.203.

[13] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are Multimodal Transformers Robust to Missing Modality?," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 18156–18165. doi: 10.1109/CVPR52688.2022.01764.

[14] J. Zeng, T. Liu, and J. Zhou, "Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2022, pp. 1545–1554. doi: 10.1145/3477495.3532064.
[15] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–14, 2023, doi: 10.1109/TPAMI.2023.3234553.
[16] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient Multimodal Transformer With Dual-Level Feature Restoration for Robust Multimodal Sentiment Analysis," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 309–325, Jan. 2024, doi: 10.1109/TAFFC.2023.3274829.
[17] R. Lin and H. Hu, "MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1686–1702, Dec. 2023, doi: 10.1162/tacl_a_00628.
[18] Z. Guo, T. Jin, and Z. Zhao, "Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1726–1736. doi: 10.18653/v1/2024.acl-long.94.
[19] P. Xiang, C. Lin, K. Wu, and O. Bai, "MultiMAE-DER: Multimodal Masked Autoencoder for Dynamic Emotion Recognition," in *2024 14th International Conference on Pattern Recognition Systems (ICPRS)*, Jul. 2024, pp. 1–7. doi: 10.1109/ICPRS62101.2024.10677820.
[20] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Robust Multimodal Learning With Missing Modalities via Parameter-Efficient Adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 2, pp. 742–754, Feb. 2025, doi: 10.1109/TPAMI.2024.3476487.
[21] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, Oct. 2020, pp. 1122–1131. doi: 10.1145/3394171.3413678.
[22] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, "Multi-modal sarcasm detection via cross-modal graph convolutional network," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 1767–1777. doi: 10.18653/v1/2022.acl-long.124.
[23] H. Liu, R. Wei, G. Tu, J. Lin, C. Liu, and D. Jiang, "Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection," *Inf. Fusion*, vol. 108, p. 102353, Aug. 2024, doi: 10.1016/j.inffus.2024.102353.
[24] S. Pramanick, A. Roy, and V. M. P. Johns, "Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2022, pp. 546–556. doi: 10.1109/WACV51458.2022.00062.
[25] Y. Gao, H. Wu, and L. Zhang, "Multi-level Conflict-Aware Network for Multi-modal Sentiment Analysis." Feb. 14, 2025. doi: 10.20944/preprints202502.1056.v1.
[26] R. Grzeczkowicz *et al.*, "Uncertainty-aware multimodal emotion recognition through dirichlet parameterization," *arXiv Prepr. arXiv2602.09121*, 2026.
[27] P. Gong, J. Liu, X. Zhang, X. Li, L. Wei, and H. He, "Towards robust sentiment analysis with multimodal interaction graph and hybrid contrastive learning," *Pattern Recognit.*, vol. 169, p. 111870, Jan. 2026, doi: 10.1016/j.patcog.2025.111870.
[28] G. Hu and others, "Recent advances in multimodal affective computing: An NLP perspective," *arXiv Prepr. arXiv2409.07388*, 2024.
[29] B. Schuller and others, "Affective computing has changed: The foundation model disruption," *arXiv Prepr. arXiv2409.08907*, 2024.
[30] Y. Zhao, Y. Peng, L. Zhang, Q. Sun, Z. Zhang, and Y. Zhuang, "Multimodal foundation model-driven user interest modeling and behavior analysis on short video platforms," *arXiv Prepr. arXiv2509.04751*, 2025.
[31] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*. Jul. 26, 2019.
[32] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - a collaborative voice analysis repository for speech technologies," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964. doi: 10.1109/ICASSP.2014.6853739.
[33] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson, "Facial expression analysis with AFFDEX and FACET: A validation study," *Behav. Res. Methods*, vol. 50, no. 4, pp. 1446–1460, Aug. 2018, doi: 10.3758/s13428-017-0996-1.