MACR: MULTIMODAL AFFECTIVE CONTRADICTION REASONING FOR EMOTION UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal emotion models can recognize individual affective cues yet misinterpret their combination when fusion suppresses the disagreement that carries meaning. We treat cross-source affective disagreement as a control signal for inference, determining whether evidence should be aggregated directly or should undergo arbitration before fusion. We introduce **MACR**, the Multimodal Affective Contradiction Reasoner, a recognize-then-arbitrate framework that preserves source-specific evidence and evaluates disagreement before fusion. MACR separates cross-source disagreement from source uncertainty: the former determines when to invoke cross-source reasoning, while the latter informs how much each source contributes. Strong disagreement activates cross-source attention before weighted fusion; low-conflict inputs follow a direct aggregation path. The resulting policy adapts how sources interact, rather than merely adjusting their contributions within a fixed fusion procedure. Across four multimodal benchmarks, MACR outperforms fine-tuned baselines, achieving F1 scores of 89.2 on HFM and 80.9 on MVSA-M. On a 480-sample, manually verified HFM diagnostic split, it improves F1 over holistic fusion by 4.7 points on conflict cases, versus 0.7 on non-conflict cases. These results support a broader principle for multimodal inference: preserve disagreement long enough for it to determine how evidence should be reconciled.

1 INTRODUCTION

Emotional meaning can reside in the disagreement between modalities, even when each individual cue is unambiguous. Positive words paired with a dismissive expression may convey sarcasm precisely because their literal and visible affect diverge. A model can therefore recognize both cues correctly and still misinterpret the emotion they express together. Advances in multimodal large language models (MLLMs) have strengthened perception and general reasoning (Alayrac et al., 2022; Liu et al., 2023), but have not resolved this gap between recognizing affective evidence and interpreting its relationships (Yang et al., 2024; Zhang et al., 2023).

A fusion-first formulation asks how to combine sources into a useful representation, leaving the significance of disagreement implicit. Yet conflicting cues may reflect either unreliable evidence or an informative contrast; treating both as variation to suppress can erase the distinction needed for interpretation. Figure 1 illustrates the resulting failure: an intermediate prediction can suppress the contrast that carries the intended affect. In our HFM error analysis, cross-modal affective conflict is the largest attributed error category, accounting for 26 of the 51 errors made by a holistic baseline (Sec. 4.3).

We treat cross-source affective disagreement as a *control signal for inference*. The essential distinction is between cross-source disagreement and source uncertainty: two confident sources can disagree, while an uncertain source need not contradict the others. Cross-source disagreement determines when to arbitrate; uncertainty informs how much to trust each source. We call this *multimodal affective contradiction reasoning*: preserving source-specific evidence, estimating potentially conflicting cues, and conditioning their interaction before committing to a joint judgment. The question shifts from how much each modality should contribute to when its relationship with other modalities requires further reasoning.

We instantiate this perspective with **MACR**, the *Multimodal Affective Contradiction Reasoner*. MACR retains each source's affective evidence in a private subspace, while measuring disagreement

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

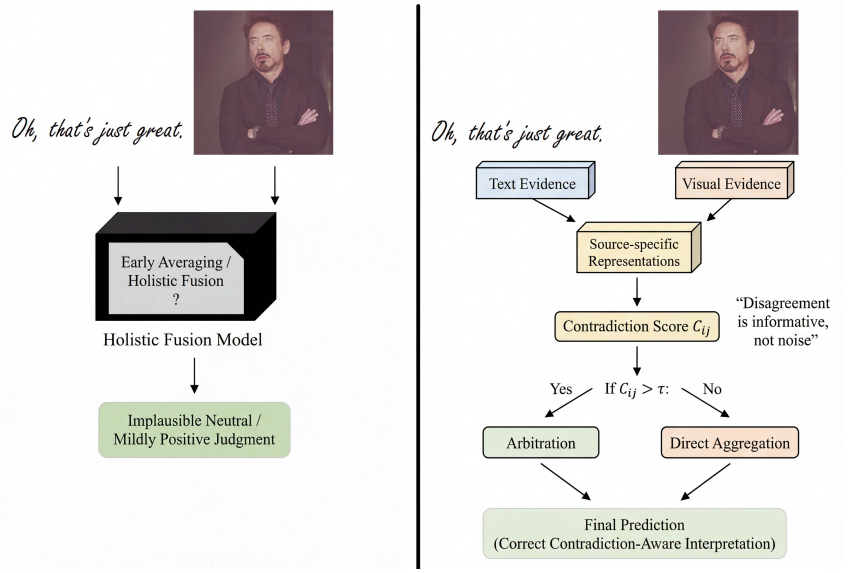


Figure 1: **Disagreement changes how evidence should be combined.** In this illustrative sarcasm example, holistic fusion can suppress the contrast between positive wording and negative expression, obscuring the relationship between literal wording and visual affect. MACR preserves source-specific evidence, measures disagreement before fusion, and invokes arbitration under strong disagreement. Low-conflict evidence follows direct aggregation: the relationship between sources determines their interaction before the final prediction.

in the shared embedding space before projection. This ordering avoids measuring contradiction through the geometric separation introduced to preserve source identity. Strong disagreement activates cross-source attention; low-conflict inputs proceed directly to aggregation. In both paths, source weights incorporate disagreement and uncertainty. The resulting *recognize-then-arbitrate* policy adapts cross-source interaction to the evidence available for each prediction.

We evaluate MACR on four multimodal benchmarks, where it outperforms task-fine-tuned baselines and reaches F1 scores of 89.2 on HFM and 80.9 on MVSA-M. On a held-out, manually verified subset of 480 HFM samples, it improves F1 over holistic fusion by 4.7 points on conflict cases versus 0.7 on non-conflict cases. Direct scorer diagnostics against human labels and matched-control ablations provide complementary evidence for contradiction-sensitive routing. These findings motivate multimodal learners whose inference policies adapt to relationships between sources, alongside representations that capture the evidence within each source.

Our contributions are threefold:

- **Problem formulation.** We frame affective contradiction reasoning as a distinct inference problem, separating the recognition of individual cues from the reconciliation of their potentially incompatible meanings.
- **Conditional arbitration.** We introduce MACR, which preserves source-specific evidence and uses contradiction to gate cross-source reasoning, with uncertainty informing source weighting.
- **Diagnostic evidence.** We combine four-benchmark evaluation with a human-verified conflict split, direct scorer diagnostics, and matched-control ablations to test where arbitration helps and what drives its gains.

2 Related Work

2.1 MLLMs for Emotion Understanding

Multimodal Large Language Models combine strong visual encoders with large language models for broad perception and reasoning (Alayrac et al., 2022; Liu et al., 2023; Tong et al., 2024), with

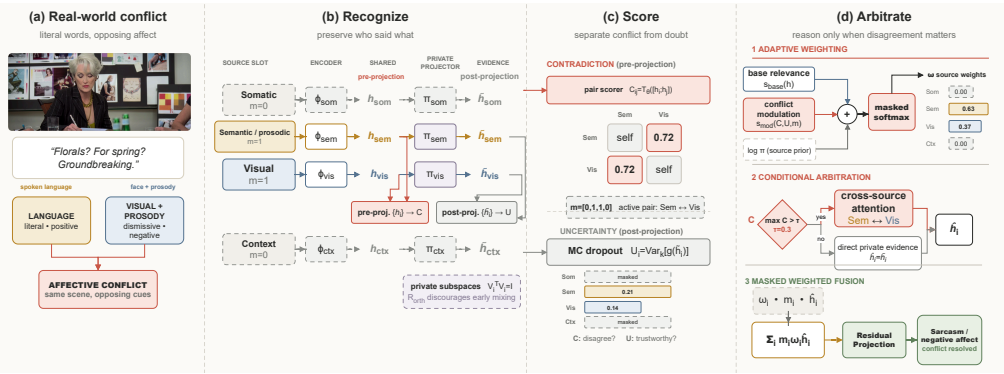


Figure 2: **MACR: Recognize, Score, and Arbitrate.** Source-specific projections retain affective evidence, while pre-projection embeddings provide shared coordinates for contradiction scoring. Strong disagreement activates cross-source attention; otherwise, evidence proceeds directly to aggregation. In both paths, source relevance, contradiction, and uncertainty jointly determine the masked aggregation weights.

general multimodal competence stress-tested by recent benchmarks (Fu et al., 2025; Liu et al., 2025; Li et al., 2024; Yue et al., 2024; Lu et al., 2024). Yet affective performance remains noticeably weaker than on general vision-language benchmarks (Yang et al., 2024; Zhang et al., 2023), and prior work attributes this to weak cross-modal attention, language over-reliance, or distraction from redundant visual content (Zhang et al., 2025). These observations mostly describe failures of perception or alignment. Our focus is narrower: even when the relevant cues are present, models can still fail because they do not reason about *disagreement* between affective signals.

2.2 Conflict, Sarcasm, and Multimodal Disagreement

Affective conflict is especially visible in sarcasm and multimodal sentiment, where text, facial expression, scene context, or prosody may point in different directions (Liu et al., 2022; Niu et al., 2016; Yang et al., 2021), and multimodal emotion recognition is widely viewed as depending on structured cross-source interaction rather than single-stream perception (Zhao et al., 2021). Existing approaches improve robustness through stronger features, specialized encoders, or task-specific fine-tuning (Cheng et al., 2024; Xie et al., 2024; Yang et al., 2021; Mai et al., 2023), but still treat disagreement as something to be fused away within a single representation. This makes them vulnerable when the contradiction itself determines the correct interpretation.

2.3 Reasoning-Oriented and Training-Free Approaches

Recent work explores both fine-tuned and training-free strategies (Zhang et al., 2025; Zhu et al., 2024; Bi et al., 2025): fine-tuned systems inject emotion-specific supervision or components (Cheng et al., 2024; Xie et al., 2024), while training-free methods improve inference through prompting, token selection, or attention redesign (Zhang et al., 2025). These improve efficiency or perception quality, but rarely isolate *contradiction reasoning* as the central failure mode. Our work is closer to a diagnosis-first perspective: we treat contradiction-heavy examples as the key stress test and use a lightweight recognize-then-arbitrate design to study how preserving modality-specific evidence changes behavior.

3 Methodology

MACR separates two decisions in multimodal inference: *whether sources need to interact* and *how their evidence should be weighted*. Contradiction controls the first decision and informs the second; uncertainty conditions the allocation of trust. The resulting recognize–score–arbitrate policy rests on a distinction between representational roles: sources are compared in shared coordinates, but their evidence remains source-specific until the interaction policy has been determined.

Problem setup and objective. Let $\mathcal{X} = \{\mathbf{x}_m\}_{m\in\mathcal{M}}$ be a multimodal observation with target affect label y, where $\mathcal{M}$ indexes the input modalities. MACR maintains four fixed source slots, $\mathcal{S} = \{1, 2, 3, 4\}$, corresponding to somatic, semantic/prosodic, visual, and contextual evidence. For each sample, the nonempty set $\mathcal{I} \subseteq \mathcal{S}$ contains the available sources, and $m_i = 1$ exactly when $i \in \mathcal{I}$. Unavailable slots are zero-filled and hard-masked, preserving a fixed routing interface across different modality configurations.

Let Θ denote the trainable parameters and $p_\Theta(y \mid \mathcal{X})$ the predictive distribution. Training minimizes

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{orth}}\mathcal{R}_{\text{orth}} + \lambda_{\text{contra}}\mathcal{L}_T, \tag{1}$$

under the projection constraints defined below. Here $\mathcal{L}_{\text{task}} = \mathbb{E}_{(\mathcal{X},y)\sim\mathcal{D}}[-\log p_\Theta(y \mid \mathcal{X})]$ is cross-entropy over the empirical training distribution $\mathcal{D}$, $\mathcal{R}_{\text{orth}}$ penalizes cross-subspace overlap, and $\mathcal{L}_T$ supervises contradiction scoring. The nonnegative coefficients λ_{orth} and λ_{contra} balance these objectives.

3.1 Recognize: Source-Specific Affective Evidence

For each available source i, an encoder maps its assigned observations $\mathcal{X}_i$ to an affect-aware representation $\mathbf{h}_i = \phi_i(\mathcal{X}_i) \in \mathbb{R}^d$. These representations share an embedding space for comparison. To retain a separate affective component for each source, we introduce a source-specific projection:

$$\tilde{\mathbf{h}}_i = \mathbf{P}_i\mathbf{h}_i, \qquad \mathbf{P}_i = \mathbf{V}_i\mathbf{V}_i^\top, \qquad \mathbf{V}_i^\top\mathbf{V}_i = \mathbf{I}_{d_s}, \tag{2}$$

where $\mathbf{V}_i \in \mathbb{R}^{d\times d_s}$ is a learnable basis and $d_s < d$. We maintain the constraint by re-orthogonalizing the basis after each optimizer step.

The constraint gives the projection a precise geometric meaning: $\mathbf{P}_i^\top = \mathbf{P}_i$ and $\mathbf{P}_i^2 = \mathbf{P}_i$. For a fixed basis, $\tilde{\mathbf{h}}_i$ is therefore the closest representation to $\mathbf{h}_i$ within $\text{range}(\mathbf{V}_i)$ in Euclidean distance. To discourage retained evidence from occupying overlapping source subspaces, we use

$$\mathcal{R}_{\text{orth}} = \sum_{i<j} \mathbb{E}_{\mathcal{X}}\left[\left\|\mathbf{P}_j\tilde{\mathbf{h}}_i\right\|_2^2\right]. \tag{3}$$

Because $\mathbf{V}_j$ has orthonormal columns, each term equals $\|\mathbf{V}_j^\top\tilde{\mathbf{h}}_i\|_2^2$: the energy retained for source i that is also visible in source j's subspace. The task loss and this data-dependent penalty jointly balance predictive information against cross-subspace overlap.

This construction retains source identity without requiring the projected representations to serve as a common metric space. That distinction determines where contradiction must be measured.

3.2 Score Contradiction: Disagreement and Uncertainty

Contradiction in shared coordinates. Different projections can separate even identical evidence: $\mathbf{h}_i = \mathbf{h}_j$ does not imply $\mathbf{P}_i\mathbf{h}_i = \mathbf{P}_j\mathbf{h}_j$. A discrepancy measured after projection can therefore reflect the geometry of source separation rather than a difference in affective content. MACR instead scores contradiction using the shared, pre-projection representations.

A distributional reference for this comparison is

$$\mathcal{C}^\star_{ij} = D_{\text{JS}}(p_i\|p_j), \qquad p_i = \mathcal{N}\big(\mathbf{h}_i, \sigma^2\mathbf{I}_d\big), \tag{4}$$

where D_{JS} is the Jensen–Shannon divergence and $\sigma > 0$ specifies a common reference scale. The shared covariance makes this an idealized comparison under the same perturbation model for each source.

The operational disagreement score is learned:

$$\mathcal{C}_{ij} = \sigma\big(T_\theta\big([\mathbf{h}_i; \mathbf{h}_j]\big)\big), \tag{5}$$

where $[\cdot;\cdot]$ denotes concatenation, T_θ is a lightweight multilayer perceptron, and $\sigma(\cdot)$ denotes the sigmoid function. Its supervision, $\mathcal{L}_T$, combines pseudo contradiction targets from frozen unimodal affect heads with a pairwise ranking loss (Appendix B). Equation (4) motivates comparison in shared coordinates; affective pseudo-supervision defines the disagreement learned by T_θ.

Uncertainty within each source. Contradiction concerns a relationship between sources; uncertainty concerns the stability of an individual source. We estimate a source-level uncertainty proxy from Monte Carlo dropout:

$$\bar{\tilde{\mathbf{h}}}_i = \frac{1}{K}\sum_{k=1}^{K} \tilde{\mathbf{h}}_i^{(k)}, \qquad \mathcal{U}_i = \frac{1}{K}\sum_{k=1}^{K} \left\| \tilde{\mathbf{h}}_i^{(k)} - \bar{\tilde{\mathbf{h}}}_i \right\|_2^2, \tag{6}$$

where $\tilde{\mathbf{h}}_i^{(k)}$ is the projected representation from the k-th stochastic pass and K is the number of passes. This dispersion is the trace of the empirical covariance of the stochastic representations.

The contradiction scores $\mathbf{C} = \{\mathcal{C}_{ij}\}$ and uncertainty estimates $\mathbf{U} = (\mathcal{U}_i)_{i=1}^4$ describe different properties of the evidence. Two stable sources can disagree strongly, while an unstable source need not contradict the others. MACR consequently uses contradiction to trigger interaction and both signals to condition source weighting.

3.3 Arbitrate: Reason Only When Disagreement Matters

Conditional interaction. Let $c(\mathcal{X})$ be the maximum contradiction score over available source pairs, $c(\mathcal{X}) = \max_{i,j\in\mathcal{I},\, i\neq j} \mathcal{C}_{ij}$. When only one source is available, we set $\mathbf{C} = \mathbf{0}$ and $c(\mathcal{X}) = 0$. Writing $\tilde{\mathbf{H}} = [\tilde{\mathbf{h}}_1; \ldots; \tilde{\mathbf{h}}_4]$, the representation passed to aggregation is

$$\hat{\mathbf{h}}_i = \begin{cases} \mathcal{A}_i(\tilde{\mathbf{H}}, \mathbf{m}), & c(\mathcal{X}) > \tau, \\ \tilde{\mathbf{h}}_i, & c(\mathcal{X}) \le \tau, \end{cases} \tag{7}$$

where $\mathcal{A}_i$ denotes the cross-source attention refinement for source i, with unavailable sources masked. We select $\tau = 0.3$ on the validation split and keep it fixed across datasets unless stated otherwise. Cross-source attention is executed only on the high-disagreement branch; both branches use the same allocation rule below.

Source allocation. Source weights depend on instance relevance and the current configuration of disagreement, uncertainty, and availability:

$$\mathbf{s}_{\text{base}} = W_b[\mathbf{h}_1; \mathbf{h}_2; \mathbf{h}_3; \mathbf{h}_4] + b_b, \qquad \mathbf{s}_{\text{mod}} = W_m[\text{vec}(\mathbf{C}); \mathbf{U}; \mathbf{m}] + b_m. \tag{8}$$

Here vec flattens the contradiction tensor, and W_b, W_m, b_b, b_m are learned parameters. The base logits encode instance-level source relevance; the modulation logits adjust that relevance using the relationships and uncertainty of the observed evidence. Let $s_i = s_{\text{base},i} + s_{\text{mod},i}$ and retain the fixed source prior $\boldsymbol{\pi} = (0.2, 0.3, 0.3, 0.2)$ in the slot order defined above.

For fixed logits, the weighting rule admits an equivalent variational characterization. Let $\Delta_{\mathcal{I}}$ be the set of probability vectors whose support is restricted to the available sources, and let $\pi_i^{\mathcal{I}} = \pi_i / \sum_{j\in\mathcal{I}} \pi_j$ be the prior normalized over that support. Then

$$\boldsymbol{\omega} = \arg\max_{\mathbf{w}\in\Delta_{\mathcal{I}}} \left\{ \sum_{i\in\mathcal{I}} w_i s_i - D_{\text{KL}}\big(\mathbf{w} \| \boldsymbol{\pi}^{\mathcal{I}}\big) \right\}, \tag{9}$$

where $D_{\text{KL}}(\mathbf{w}\|\boldsymbol{\pi}^{\mathcal{I}}) = \sum_{i\in\mathcal{I}} w_i \log(w_i/\pi_i^{\mathcal{I}})$ is the Kullback–Leibler divergence. The objective trades off the learned source scores against departure from the reference allocation. Its unique maximizer follows from $\omega_i \propto \pi_i^{\mathcal{I}} \exp(s_i)$:

$$\omega_i = \begin{cases} \dfrac{\pi_i \exp(s_{\text{base},i} + s_{\text{mod},i})}{\sum_{j\in\mathcal{I}} \pi_j \exp(s_{\text{base},j} + s_{\text{mod},j})}, & i \in \mathcal{I}, \\ 0, & i \notin \mathcal{I}. \end{cases} \tag{10}$$

This is the masked softmax of $\log \pi_i + s_{\text{base},i} + s_{\text{mod},i}$. Equal learned logits recover the normalized prior; differences in the logits adjust the relative source weights. Missing sources receive exactly zero mass, and normalization is performed only over observed evidence.

Finally, MACR forms the task representation

$$\mathbf{z} = \mathrm{ResidualProj}\left(\sum_{i=1}^{4} m_i \omega_i \hat{\mathbf{h}}_i\right), \tag{11}$$

which is passed to the task classifier. For a single available source, arbitration is disabled and its normalized weight is one, making single-source prediction a boundary case of the same inference rule.

The resulting mechanism separates interaction from allocation: contradiction determines whether evidence requires cross-source refinement, while relevance, contradiction, and uncertainty determine how the resulting representations contribute to the prediction.

4 Experiments

4.1 Experimental Setup

Datasets. Our primary evaluation covers four multimodal sentiment benchmarks: MVSA-S and MVSA-M (Niu et al., 2016), TumEmo (Yang et al., 2021), and HFM (Cai et al., 2019; Liu et al., 2022). A manually verified HFM subset provides conflict/non-conflict labels for source-level diagnosis. We additionally evaluate visual emotion recognition on Emotion6 (Peng et al., 2015), EmoSet (Yang et al., 2023), WebEmo7/25, and Abstract (Machajdik & Hanbury, 2010), and physiological-multimodal affect understanding on ASCERTAIN (Subramanian et al., 2018). Results for these additional settings appear in Appendix H.

Implementation. MACR uses Llama-3-8B and VILA-8B-style backbones (Dubey et al., 2024; Lin et al., 2024). Text and images/video frames use the backbone text and vision encoders; audio/prosody uses a lightweight conformer head; and physiological streams (EEG/ECG/GSR) use a two-layer temporal MLP. All sources are mapped into the shared d-dimensional space. Cross-source arbitration uses eight-head attention, with $\tau = 0.3$ selected on the task validation split from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. We train with AdamW, a learning rate of 2×10^{-5}, weight decay of 0.01, and mixed precision on $8\times$A100 GPUs. The four source slots use the prior $[0.2, 0.3, 0.3, 0.2]$; unavailable sources are zero-filled and hard-masked before weight normalization. Appendix C specifies the modality-to-source mapping and active slots for each benchmark.

Baselines and reporting. Table 1 groups methods by training regime. Training-free MLLMs are evaluated out of the box, while M^{2}CL, MULSER, CMGCN, SPFVTE, and MODA† are trained on the same task data as MACR. Matched controls use the same backbone, training data, and optimization budget while changing the arbitration mechanism (Table 4). MACR is evaluated over three random seeds using standard train/validation/test splits. We report mean $\pm$ standard deviation for MVSA-M, TumEmo, and HFM in the main comparison, and for MVSA-M and HFM in the component ablations.

4.2 Benchmark Performance and Conflict Diagnosis

Multimodal sentiment. MACR achieves strong performance across all four benchmarks in Table 1, reaching 89.2 F1 on HFM and 80.9 on MVSA-M. Relative to the task-fine-tuned MODA† baseline, F1 improves by 1.4 points on MVSA-S, 1.9 on MVSA-M, and 1.1 on both TumEmo and HFM. We examine the source of these gains through a conflict-focused analysis of HFM's sarcasm-rich examples.

Gains concentrate on conflicting evidence. We construct a 480-sample HFM diagnostic split by sampling from both high- and low-disagreement regions of frozen text-side and vision-side affect predictors, then manually verifying the conflict/non-conflict labels. The scorer learns from pseudo targets on the original training split; the verified labels are used for held-out evaluation. MACR raises conflict-subset F1 from 81.2 to 85.9, a gain of 4.7 points, compared with 0.7 points on non-conflict samples (Table 2). The 95% bootstrap confidence interval for the conflict-subset gain is $[+2.5, +6.9]$ across 10,000 resamples. The improvement is thus concentrated where the final judgment requires reconciling divergent affective evidence.

Table 1: **Multimodal sentiment results.** Accuracy and F1 are reported on a 0–100 scale. Methods are grouped by training regime; † denotes the fine-tuned MODA variant. Best and second-best results are bold and underlined.

Method	MVSA-S		MVSA-M		TumEmo		HFM	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Training-free: proprietary MLLMs								
GPT-4V	50.7	57.0	60.9	63.1	60.8	61.2	76.4	76.5
Gemini 1.0 Pro	63.4	63.7	69.9	65.7	59.8	58.2	67.4	68.3
Claude 3 Opus	62.6	61.3	63.5	62.9	58.0	57.4	67.9	68.7
QWen-VL-Max	64.7	64.5	66.9	62.7	56.5	59.5	69.6	70.1
Training-free: open-source MLLMs								
Mini-Gemini-HD	42.3	57.1	48.7	64.3	24.6	39.5	49.8	59.3
LLaVA-NeXT	59.1	59.3	61.7	60.7	54.7	53.3	57.2	55.1
Cambrian-1	69.4	66.1	66.2	57.9	43.9	34.4	51.2	48.7
MODA	70.2	70.5	62.8	61.9	55.9	54.8	58.5	56.3
Fine-tuned on task data								
M^2CL	75.5	74.2	73.2	70.5	68.8	68.7	–	–
MULSER	75.7	75.5	73.9	73.8	77.5	77.5	–	–
CMGCN	73.3	72.0	69.7	68.3	–	–	87.5	84.1
SPFVTE	80.6	80.1	79.9	78.8	–	–	88.3	87.9
MODA†	<u>81.0</u>	<u>80.3</u>	<u>80.2</u>	<u>79.0</u>	<u>77.8</u>	<u>77.8</u>	<u>88.5</u>	<u>88.1</u>
MACR (Ours)	**82.3**	**81.7**	$\mathbf{81.6} \pm 0.3$	$\mathbf{80.9} \pm 0.5$	$\mathbf{79.3} \pm 0.4$	$\mathbf{78.9} \pm 0.5$	$\mathbf{89.7} \pm 0.3$	$\mathbf{89.2} \pm 0.4$

MACR results on MVSA-M, TumEmo, and HFM are mean ± standard deviation over three seeds.

Table 2: **Conflict-focused evaluation on 480 HFM samples.** F1 gains are measured relative to MACR (Holistic) in absolute points.

Subset	Holistic F1	MACR F1	Gain
Overall	86.6	89.5	+2.9
Conflict	81.2	85.9	**+4.7**
Non-conflict	90.5	91.2	+0.7

Paired McNemar test on the diagnostic split: $\chi^2 = 6.21$, $p = 0.013$.

4.3 Error Decomposition and Counterfactual Swap Test

Error decomposition. On the 494-sample HFM test set, the holistic baseline makes 51 errors. Manual attribution under the conflict annotation protocol assigns 26 errors to cross-modal affective conflict, 15 to weak or missing perceptual cues, and 10 to label ambiguity. Conflict accounts for 51% of the attributed errors, making it the largest category (Appendix H). This diagnosis connects the conflict-specific gains to a prominent failure mode of holistic inference.

Counterfactual modality swaps. We construct $N = 100$ paired examples from the HFM test set. For each conflict-prone example, we retain the text and replace its image with one whose affective polarity opposes the text. Under this replacement, the holistic model's error rate rises from 13.0% to 17.8%, while MACR's rises from 10.3% to 13.1%: an increase of 4.8 versus 2.8 percentage points. MACR also has a lower prediction flip rate, 15.4% versus 21.3%, where a flip denotes a change in the predicted label after replacement. Under the same visual perturbation, MACR exhibits both a smaller error increase and fewer prediction changes. Appendix H reports the complete swap results.

4.4 Contradiction Scorer Diagnostics

Conflict detection. Table 3 evaluates the contradiction scorer against the human conflict labels on the 480-sample HFM split. The scorer reaches an AUROC of 0.79 and an F1 of 0.76 at the validation-selected threshold $\tau = 0.3$. Training the same-capacity scorer on random labels yields an

Table 3: **Contradiction scorer diagnostics on 480 HFM samples.** Detection metrics and human-label agreement are shown in separate panels. Threshold-dependent detection metrics in panel (a) use the validation-selected routing threshold $\tau = 0.3$; panel (b) reports a separate human-label agreement analysis.

(a) Conflict detection

Metric	Value
AUROC	0.79
AUROC (random labels)	0.51
F1	0.76
Precision	0.79
Recall	0.74
Trigger rate	15.0%

(b) Human-label agreement

Band / statistic	Value
High ($\mathcal{C} > 0.7$)	87.2%
Middle ($0.5 \leq \mathcal{C} \leq 0.7$)	76.5%
Low ($\mathcal{C} < 0.5$)	55.9%
Overall agreement	78.6%
Scorer–human κ	0.52
Annotator Fleiss' κ	0.74

Table 4: **Scoring and interaction under matched training conditions.** All variants share the backbone, training data, and optimization budget. F1 is reported on a 0–100 scale.

Variant	MVSA-M F1	HFM F1
No scorer (always-on attention)	78.0	85.0
No attention (scoring only)	79.6	86.5
MACR (Full)	**80.9**	**89.2**

AUROC of 0.51, compared with 0.79 under affective pseudo targets. The pseudo-label construction is detailed in Appendix B.

Agreement across score bands. Agreement with human labels increases from 55.9% in the low-score band to 87.2% in the high-score band. Overall agreement is 78.6%, with scorer–human $\kappa = 0.52$; the three annotators have Fleiss' $\kappa = 0.74$. The score-stratified results characterize how the routing signal corresponds to human judgments across different levels of disagreement.

4.5 Ablation Studies

Scoring and interaction. Table 4 separates the roles of contradiction scoring and cross-source attention under the same backbone, data, and optimization budget. The *no-scorer* variant removes the contradiction MLP and applies cross-source attention to every sample; *no-attention* retains the scorer and learned weights but bypasses cross-source attention. Full MACR outperforms both variants on both multimodal benchmarks. On HFM, F1 increases from 85.0 with always-on attention to 86.5 with scoring alone, and reaches 89.2 with the complete policy. Scoring and cross-source interaction are most effective when used together.

Contribution of each mechanism. Each component contributes across the two evaluated multimodal benchmarks (Table 5). On HFM, removing source decoupling reduces F1 by 4.4 points, removing dynamic fusion by 3.0 points, and removing conflict resolution by 2.1 points. Replacing the attention mechanism with MODA attention or SEPM VTA reduces HFM F1 to 85.7 and 84.7, respectively. These results link the gains to the joint use of source-specific representations, adaptive weighting, and contradiction-aware interaction.

4.6 Qualitative Analysis and Efficiency

Tracing an arbitration decision. Figure 3 follows an example with conflicting textual and visual affect. The source-level probabilities of positive affect are 0.62 for text and 0.19 for the image, while the contradiction score $\mathcal{C} = 0.71 > \tau$ activates cross-source attention. Visual uncertainty is higher than textual uncertainty (0.26 versus 0.18), and the final text/image weights are $(0.68, 0.32)$. MACR predicts the sarcastic interpretation instead of an intermediate neutral judgment. This trace

Table 5: **Component ablations on multimodal benchmarks.** F1 is reported on a 0–100 scale; results are mean ± standard deviation over three seeds.

Variant	MVSA-M F1	HFM F1
MACR (Full)	**80.9** ± 0.5	**89.2** ± 0.4
w/o Decoupling	77.1 ± 0.6	84.8 ± 0.6
w/o Dynamic Fusion	78.3 ± 0.8	86.2 ± 0.5
w/o Conflict Resolution	78.9 ± 0.7	87.1 ± 0.3
Replace w/ MODA Attn	77.9 ± 0.4	85.7 ± 0.5
Replace w/ SEPM VTA	77.2 ± 0.5	84.7 ± 0.3

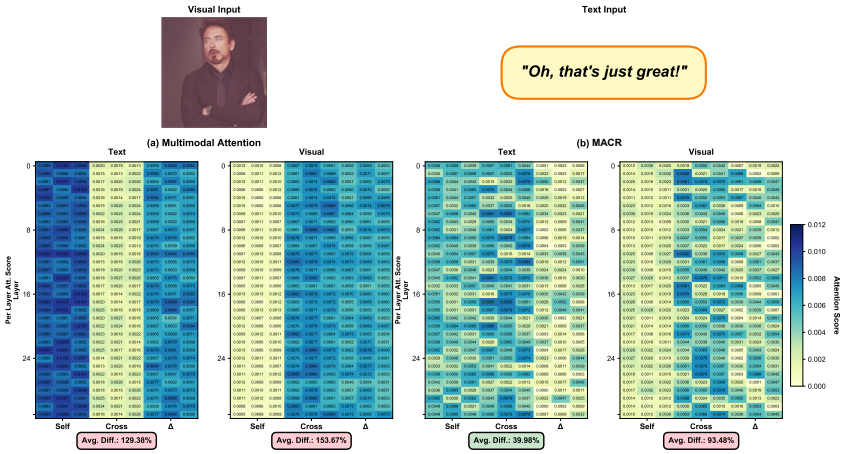


Figure 3: **Cross-source interaction in a sarcasm example.** The baseline concentrates attention within modalities; MACR redistributes interaction across sources and recovers the sarcastic interpretation.

makes the progression from source evidence to disagreement, interaction, and final weighting explicit (Appendix G). Appendix D reports the attention-difference metric Δ. The t-SNE diagnostics in Appendix H give a silhouette score of 0.412 and a 10-NN score of 67.8% for MACR, compared with 0.288 and 56.4% for holistic fusion.

Efficiency and sensitivity. MACR adds 186M parameters ($\approx 2.3\%$) and increases per-sample latency from 23 ms to 27 ms on a single A100 at batch size 1 in FP16. Cross-source attention is activated on 12% of samples on average, while the $K = 20$ Monte Carlo dropout passes are confined to the projection heads ($\sim$38M parameters each, less than 0.2 ms per pass). Performance remains stable under moderate changes to τ and the source prior. Training time and peak memory are reported in Appendix G. The MC dropout configuration is described in Appendix E.

5 Conclusion

We study *affective contradiction reasoning* as an important failure mode in multimodal emotion understanding and present **MACR**, a compact recognize-then-arbitrate framework that preserves source-specific evidence, scores contradiction, and intervenes only when disagreement is strong. The clearest gains appear in contradiction-prone settings, and the conflict-focused diagnostic split shows these gains concentrate on manually verified conflict cases. Because the paper studies human-centered affective signals, we emphasize that such systems carry privacy, bias, and misuse risks and should be treated as research tools rather than deployment-ready products.

REPRODUCIBILITY STATEMENT

We describe the full inference procedure in Section 3: the training objective in Eq. (1), the source-specific projection and its orthogonality penalty in Eqs. (2) and (3), the contradiction scorer in Eq. (5), the uncertainty proxy in Eq. (6), and the conditional arbitration and source allocation rules in Eqs. (7) and (10). Appendix C reports the modality-to-source mapping for every benchmark and the active slots used in each configuration. Algorithm 1 restates the end-to-end arbitration procedure in step-by-step form.

All optimization settings are given in Section 4: AdamW with a learning rate of 2×10^{-5} and weight decay of 0.01, mixed precision on $8\times$A100 GPUs, $K = 20$ Monte Carlo dropout passes, and the fixed source prior $\boldsymbol{\pi} = (0.2, 0.3, 0.3, 0.2)$. The threshold τ is selected on the task validation split by sweeping $\{0.1, 0.2, 0.3, 0.4, 0.5\}$; The threshold-sensitivity analysis is reported in Appendix C. The Monte Carlo dropout configuration used for uncertainty estimation is described in Appendix E. Contradiction-scorer supervision, including pseudo-target construction and the pairwise ranking term, is specified in Appendix B, and the attention-difference metric Δ is defined in Appendix D.

Reported main-comparison results for MVSA-M, TumEmo, and HFM are means over three random seeds with standard deviations; all benchmarks use their standard train/validation/test splits. Conflict-subset gains are reported with a 95% bootstrap confidence interval over 10,000 resamples, and the paired McNemar test is used for the diagnostic-split comparison. The manually verified conflict labels, the counterfactual swap protocol, the error-attribution procedure, and the additional visual-emotion and physiological settings are documented in Appendix H. Efficiency measurements are reported in Table 7.

ETHICS STATEMENT

This work studies human-centered affective signals, and the models and analyses presented here are intended strictly for research use. Affective recognition carries privacy, bias, and misuse risks, particularly in surveillance or high-stakes decision settings, and the physiological and personality-related streams in ASCERTAIN are especially sensitive. We therefore make no deployment claims, and we do not recommend the use of these methods for automated judgments about individuals. The benchmarks used in this paper are existing, publicly documented datasets; we introduce no new data collection and no new human-subject procedures beyond the manual verification of conflict labels described in Section 4. A fuller discussion of ethical considerations and of the limitations of the present evaluation appears in Appendix F.

THE USE OF LARGE LANGUAGE MODELS

Large language models were used solely for grammar checking and language polishing of manuscript text. They were not used to generate research ideas, to design the method or architecture, to run or analyze experiments, or to produce any experimental number, table entry, or figure. All research content, including the problem formulation, the method, the experimental design, and the analysis and interpretation of results, was conceived and carried out by the authors. All text that underwent language polishing was reviewed and approved by the authors, who take responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. PRISM: Self-Pruning Intrinsic Selection Method for Training-Free Multimodal Data Selection. arXiv preprint arXiv:2502.12119v1, 2025. URL https://arxiv.org/abs/2502.12119v1.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2506–2515. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1239. URL https://aclanthology.org/P19-1239/.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G. Hauptmann. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110805–110853. Curran Associates, Inc., 2024. doi: 10.52202/079017-3518. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c7f43ada17acc234f568dc66da527418-Abstract-Conference.html.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783v1, 2024. URL https://arxiv.org/abs/2407.21783v1.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162549–162567, 2025. doi: 10.52202/085713-4899. URL https://papers.nips.cc/paper_files/paper/2025/hash/d79a27cf2772fe00be7f341efc0eb517-Abstract-Datasets_and_Benchmarks_Track.html.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, June 2024. doi: 10.1109/CVPR52733.2024.01263. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_SEED-Bench_Benchmarking_Multimodal_Large_Language_Models_CVPR_2024_paper.html.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26689–26699, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_VILA_On_Pre-training_for_Visual_Language_Models_CVPR_2024_paper.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Hui Liu, Wenya Wang, and Haoliang Li. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4995–5006. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.333. URL https://aclanthology.org/2022.emnlp-main.333/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-Around Player? In *Computer Vision – ECCV 2024*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2025. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of

Foundation Models in Visual Contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/663bce02a0050c4a11f1eb8a7f1429d3-Paper-Conference.pdf.

Jana Machajdik and Allan Hanbury. Affective Image Classification Using Features Inspired by Psychology and Art Theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 83–92. ACM, 2010. doi: 10.1145/1873951.1873965. URL https://doi.org/10.1145/1873951.1873965.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, 14 (3):2276–2289, 2023. doi: 10.1109/taffc.2022.3172360. URL https://doi.org/10.1109/taffc.2022.3172360.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment Analysis on Multi-View Social Data. In *MultiMedia Modeling*, volume 9517 of *Lecture Notes in Computer Science*, pp. 15–27. Springer, 2016. doi: 10.1007/978-3-319-27674-8_2. URL https://doi.org/10.1007/978-3-319-27674-8_2.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher. A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 860–868, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Peng_A_Mixed_Bag_2015_CVPR_paper.html.

Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2018. doi: 10.1109/taffc.2016.2625250. URL https://doi.org/10.1109/taffc.2016.2625250.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356. Curran Associates, Inc., 2024. doi: 10.52202/079017-2771. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26596–26605, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xie_EmoVIT_Revolutionizing_Emotion_Insights_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. EmoSet: A Large-scale Visual Emotion Dataset with Rich Attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20383–20394, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yang_EmoSet_A_Large-scale_Visual_Emotion_Dataset_with_Rich_Attributes_ICCV_2023_paper.html.

Qu Yang, Mang Ye, and Bo Du. EmoLLM: Multimodal Emotional Understanding Meets Large Language Models. arXiv preprint arXiv:2406.16442, 2024. URL https://arxiv.org/abs/2406.16442.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. *IEEE Transactions on Multimedia*, 23:4014–4026, 2021. doi: 10.1109/tmm.2020.3035277. URL https://doi.org/10.1109/tmm.2020.3035277.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly Supervised Video Emotion Detection and Prediction via Cross-Modal Temporal Erasing Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18888–18897, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Weakly_Supervised_Video_Emotion_Detection_and_Prediction_via_Cross-Modal_Temporal_CVPR_2023_paper.html.

Zhicheng Zhang, Wuyou Xia, Chenxi Zhao, Zhou Yan, Xiaoqiang Liu, Yongjie Zhu, Wenyu Qin, Pengfei Wan, Di Zhang, and Jufeng Yang. MODA: MOdular Duplex Attention for Multimodal Perception, Cognition, and Emotion Understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76385–76396. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25cg.html.

Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion Recognition from Multiple Modalities: Fundamentals and Methodologies. *IEEE Signal Processing Magazine*, 38(6): 59–73, 2021. doi: 10.1109/msp.2021.3106895. URL https://doi.org/10.1109/msp.2021.3106895.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling Cross-Modality Knowledge Conflicts in Large Vision-Language Models. arXiv preprint arXiv:2410.03659, 2024. URL https://arxiv.org/abs/2410.03659.

A SOURCE-SPECIFIC SUBSPACE PARAMETERIZATION

The separation component in MACR parameterizes each evidence source as a learnable linear subspace within the shared embedding space. We want to be explicit about its status: it is a lightweight, *pragmatic mechanism* that keeps source identity inspectable until arbitration, not an architectural claim about the psychological separability of affect. One can view it as a structured bottleneck that makes the recognize-then-arbitrate policy implementable in practice.

Specifically, each source subspace $\mathcal{S}_i$ is defined by a basis matrix $\mathbf{V}_i \in \mathbb{R}^{d \times d_s}$ with $d_s < d$, constrained to be semi-orthonormal, $\mathbf{V}_i^T \mathbf{V}_i = \mathbf{I}_{d_s}$, so that $\mathbf{P}_i$ is an orthogonal projection rather than an unconstrained low-rank map. The projection operator is

$$\tilde{\mathbf{h}}_i = \mathbf{P}_i(\mathbf{h}_i) = \mathbf{V}_i \mathbf{V}_i^T \mathbf{h}_i.$$

The training objective encourages different evidence sources to remain separable through a regularization term on inter-source projections:

$$\mathcal{R}_{\text{orth}} = \sum_{i<j} \mathbb{E}_{\mathbf{x}} \left[\left\| \mathbf{V}_j \mathbf{V}_j^T \tilde{\mathbf{h}}_i \right\|_F^2 \right].$$

Empirically, this separation helps prevent premature mixing of contradictory evidence before arbitration.

The important caveat is what this mechanism does *not* claim. It does not assume that each source corresponds to an isolated psychological factor, nor that every emotionally relevant cue must be mapped to a distinct subspace. Its role is operational: it provides each source a private representational buffer so that contradiction can be measured from preserved source evidence, rather than reconstructed after fusion has already erased the disagreement. Whether psychologically motivated separation is additionally needed is a separate question that the present method does not take a position on.

A.1 QUALITATIVE FAILURE CASE ANALYSIS

While MACR improves contradiction-aware emotion understanding, several failure modes remain.

1. Missed contradiction. In some cases, genuine disagreement is too subtle to trigger arbitration. Deadpan sarcasm with only weak expressive cues may yield a low contradiction score, causing the model to fall back to the literal interpretation.

2. Incorrect arbitration. In other cases, contradiction is detected but resolved incorrectly. This commonly occurs in benign dissonance such as teasing or playful banter, where negative literal wording coexists with positive social cues. The model may identify disagreement yet still assign too much weight to the wrong source.

3. Ambiguous labels. Fine-grained emotion datasets often contain inherently ambiguous affective states. In these cases, MACR may produce a plausible alternative interpretation by exposing disagreement between sources, but that interpretation can still be counted as an error under a single-label ground truth.

B CONTRADICTION SCORE ESTIMATION

The contradiction scorer T_θ is a lightweight MLP that estimates *instance-level affective disagreement* between two sources. We train it with a combination of pseudo-label supervision and pairwise ranking, rather than simple source-identity discrimination.

Pseudo-label construction. Directly comparing raw source-specific label spaces would indeed be ill-posed, because different modalities may expose different native prediction heads. MACR therefore does *not* compare those native labels directly. Instead, each source-specific head is mapped into a shared *contradiction space* $\mathcal{A}$ before pseudo-label construction. Concretely, for each active source i, we define

$$\bar{\mathbf{p}}_i = g_i(f_i(\mathbf{x}_i)) \in \Delta^{|\mathcal{A}|-1},$$

where f_i is the frozen source-specific affect head and g_i is a fixed source-to-shared mapping. In our experiments, $\mathcal{A}$ is a coarse valence-oriented space with three bins,

$$\mathcal{A} = \{\text{negative},\ \text{neutral/ambiguous},\ \text{positive}\},$$

which is expressive enough for contradiction detection while remaining comparable across modalities. For sentiment benchmarks, g_i maps directly from source-level sentiment predictions into these three bins. For multimodal emotion settings whose native labels are finer grained, g_i collapses source predictions to the corresponding coarse valence bucket; low-confidence or mixed-affect cases are mapped to the neutral/ambiguous bin. For physiological settings such as ASCERTAIN, auxiliary source heads are calibrated to the same coarse valence space for contradiction supervision, even though the downstream task still predicts the original benchmark targets.

For each training sample and each active source pair (i, j), we then define a pseudo contradiction target:

$$y^*_{ij} = \mathbb{1}\big[\operatorname{argmax} \bar{\mathbf{p}}_i \neq \operatorname{argmax} \bar{\mathbf{p}}_j\big],$$

indicating whether source i and source j disagree on the predicted shared affect label for the *same sample*. This construction makes the supervision comparable across modalities by design: disagreement is measured only after modality-specific outputs have been normalized into the common contradiction space, so the scorer learns instance-level contradiction rather than distributional source identity or arbitrary label-set mismatch.

Training objective. The scorer T_θ takes the concatenation $[\mathbf{h}_i; \mathbf{h}_j]$ (pre-projection representations in the shared space) and outputs a scalar contradiction estimate. It is trained with:

$$\mathcal{L}_{\text{contra}} = -\,\mathbb{E}\big[y^*_{ij} \log \sigma(T_\theta([\mathbf{h}_i; \mathbf{h}_j])) + (1 - y^*_{ij}) \log(1 - \sigma(T_\theta([\mathbf{h}_i; \mathbf{h}_j])))\big].$$

We additionally apply a pairwise ranking regularizer that encourages contradiction scores to be higher for pseudo-conflict pairs than for pseudo-consistent pairs within each mini-batch:

$$\mathcal{L}_{\text{rank}} = \mathbb{E}\big[\max(0,\ T_\theta([\mathbf{h}_i; \mathbf{h}_j]^{\text{cons}}) - T_\theta([\mathbf{h}_i; \mathbf{h}_j]^{\text{conf}}) + \delta)\big],$$

where conf and cons denote pseudo-conflicting and pseudo-consistent pairs respectively, and $\delta = 0.5$ is a margin. The final contradiction loss is $\mathcal{L}_T = \mathcal{L}_{\text{contra}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}$ with $\lambda_{\text{rank}} = 0.3$.

Relation to main-paper notation. The implementation-time score $\mathcal{C}_{ij} = \sigma(T_\theta([\mathbf{h}_i; \mathbf{h}_j]))$ is the operational disagreement signal corresponding to the shared-coordinate comparison motivated by Eq. (4).The neural estimator therefore provides a learnable, instance-conditioned contradiction signal while preserving the main paper's JS-divergence interpretation as motivation rather than as the literal training objective.

Why a coarse shared contradiction space is sufficient. The contradiction scorer is not asked to solve the full downstream label space on its own. Its role is only to decide whether two sources support *compatible* or *incompatible* affective hypotheses. A coarse shared space is therefore preferable to a high-cardinality label ontology: it avoids spurious non-comparability across modalities while still preserving the sign and ambiguity information needed for arbitration. Fine-grained downstream prediction remains the responsibility of the task head after source arbitration.

Figure 4 provides the distributional view behind the operating-point metrics in the main text. The score histograms are not perfectly separated, which is expected for affective contradiction data with ambiguous and weakly expressed cases, but the shift between conflict and non-conflict subsets is large enough to support contradiction-triggered arbitration as a meaningful routing signal rather than a trivial artifact of label imbalance.

C Experimental Protocols

Modality-to-Source Mapping The recognize-then-arbitrate framework adapts to the modalities available in each dataset. Table 6 shows the mapping used in our experiments. Importantly, most datasets instantiate only a subset of evidence sources. This is why the paper's main claim should be understood as contradiction-aware arbitration over available evidence rather than as a requirement that every source be present.

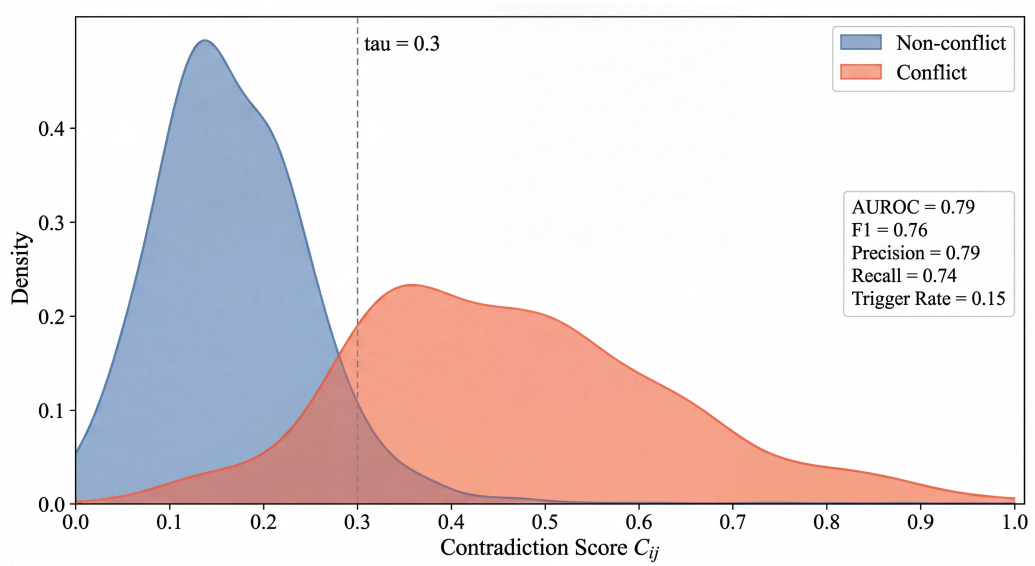


Figure 4: Distribution of contradiction scores on the manually annotated HFM diagnostic subset. Non-conflict samples concentrate near low contradiction scores, while conflict samples shift toward higher scores with realistic overlap around the operating region. The dashed line marks the validation-selected threshold $\tau = 0.3$, which yields the operating-point metrics reported in the main paper.

Table 6: Modality-to-source mapping across datasets.

Dataset	Somatic	Semantic	Visual	Contextual
MVSA-S/M, TumEmo, HFM	*Inactive*	Text	Vision (Scene/Object)	*Inactive*
Emotion6, EmoSet	*Inactive*	*Inactive*	Vision (Face/Scene)	*Inactive*
WebEmo7, WebEmo25, Abstract	*Inactive*	*Inactive*	Vision (Image-Level Semantics)	*Inactive*
ASCERTAIN	EEG, ECG, GSR	Text, Audio (Prosody)	Vision (Face/Video)	Personality / Trait History

For single-source benchmarks (Emotion6, EmoSet, WebEmo7, WebEmo25, and Abstract), pairwise contradiction is undefined; accordingly, we set $\mathbf{C} = \mathbf{0}$ and disable contradiction-triggered arbitration. In this regime, MACR no longer performs cross-source aggregation; it degenerates to a single-source projected prediction path in which the lone valid source receives unit weight after masking/renormalization, so the final representation is just that source's projected feature followed by the residual prediction head, exactly as described in Sec. 3.

Encoder Instantiation and Training Details Across datasets, text is encoded with the LLM text tower, images/video frames with the vision encoder of the backbone, audio/prosody with a lightweight conformer head, and physiological streams (EEG/ECG/GSR) with a 2-layer temporal MLP before projection into the shared d-dimensional space. Unsupported sources are hard-masked and skipped entirely. Unless stated otherwise, we tune τ on validation data from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and use $\tau = 0.3$.

How the Fixed Prior Weights Enter Arbitration The implementation uses a global source prior

$$\boldsymbol{\pi} = [0.2,\ 0.3,\ 0.3,\ 0.2],$$

ordered as *somatic*, *semantic/prosodic*, *visual*, and *contextual*. These values are *not* a separate post-hoc interpolation rule; they enter the routing network as a constant bias in logit space, exactly as written in Eq. (10) of the main paper (which shows $\log \pi_i$ explicitly):

$$\omega_i = \mathrm{softmax}_i(\log \pi_i + s_{\mathrm{base},i} + s_{\mathrm{mod},i}), \quad i \in \{1, 2, 3, 4\},$$

followed by hard masking and renormalization over the active-source set $\mathcal{I}$.

This design makes the priors conservative rather than dominant. They provide a stable default preference when contradiction evidence is weak, but they can be overridden whenever the learned

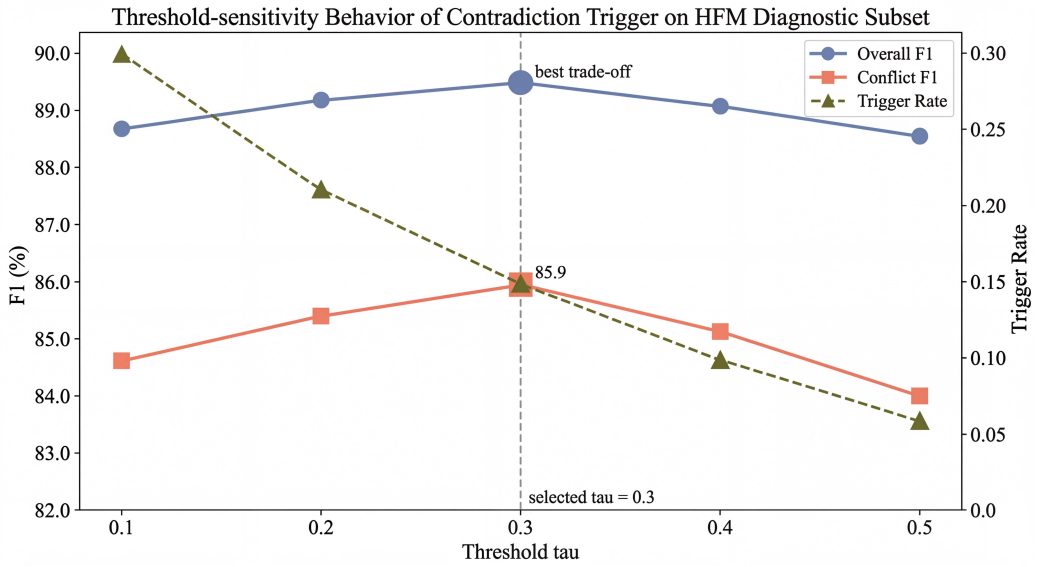


Figure 5: Threshold sensitivity of contradiction-triggered arbitration on the annotated HFM diagnostic subset. The selected operating point $\tau = 0.3$ achieves the best balance between overall diagnostic performance, conflict-case performance, and trigger frequency. Lower thresholds over-trigger arbitration, whereas higher thresholds become too conservative and miss contradiction-heavy cases.

instance-conditioned terms s_{base} and s_{mod} provide stronger evidence. The chosen values mildly favor the semantic and visual slots because those are the most frequently instantiated and most contradiction-relevant sources in the main sentiment benchmarks, while keeping somatic and contextual slots available without forcing them to dominate when present. If a source is missing, its slot is masked before softmax, so the corresponding prior contributes nothing to the final normalized weights. In principle these priors could be learned jointly with the routing network; we keep them fixed for interpretability and to preserve the ability to inspect how much of each source's weight derives from the instance-conditioned terms versus the default preference.

Data Splits and Statistical Reporting For all benchmarks we use the standard train/validation/test protocols when available. Unless otherwise stated, reported MACR results are averaged over three independent runs with different random seeds. For compactness, the main paper reports mean $\pm$ std on the contradiction-sensitive benchmarks and MACR-family ablations, while the remaining tables show the averaged point estimates only.

MLLM Baseline Prompts Zero-shot proprietary MLLMs were evaluated using task-specific prompts. These results are included as contextual reference alongside the stronger open and task-specific baselines emphasized in the main paper.

Definition of the holistic baseline The **MACR (Holistic)** variant referenced in the main paper is a same-backbone ablation introduced to isolate the effect of contradiction-aware arbitration. It keeps the same backbone encoders, shared embedding space, source-slot interface, and downstream task head as the full MACR model. However, it removes the components specific to contradiction-aware reasoning: (i) source-specific subspace separation, (ii) contradiction-score estimation and contradiction-triggered routing, and (iii) contradiction-aware cross-source attention during arbitration. Instead, all available source embeddings are merged through a single flat holistic fusion path before prediction. The comparison between **MACR (Holistic)** and **MACR (Full)** therefore isolates one design choice: whether source-specific affective evidence is preserved long enough for contradiction-aware arbitration, or instead collapsed directly through ordinary holistic fusion.

Figure 5 complements the main-paper contradiction diagnostics by showing why $\tau = 0.3$ is used throughout the experiments. Performance remains reasonably stable under moderate threshold changes, consistent with the robustness claim in Sec. 4.6, while the trigger rate decreases monotoni-

cally as the threshold becomes more conservative. The chosen point therefore represents a practical balance rather than a brittle tuning artifact.

Algorithm 1 MACR Inference

Require: Four source-slot representations $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^4$, availability mask $\mathbf{m}$, projectors $\{\pi_i\}_{i=1}^4$, threshold τ
Ensure: Fused representation $\mathbf{z}$ and arbitration metadata $\mathcal{F} = \{\boldsymbol{\omega}, \max \mathbf{C}, \mathbf{1}[\max \mathbf{C} > \tau]\}$
for $i = 1$ to 4 **do**
  $\tilde{\mathbf{h}}_i \leftarrow \pi_i(\mathbf{h}_i)$ if $m_i = 1$, else $\mathbf{0}$
end for
Estimate contradiction scores $\mathbf{C}$ from the pre-projection representations of valid source pairs (i, j) with $m_i = m_j = 1$
Estimate uncertainties $\mathcal{U}$ via Eq. (6)
Compute adaptive weights $\boldsymbol{\omega}$ via Eq. (10)
if $\max \mathbf{C} > \tau$ **then**
  $\hat{\mathbf{h}}_i \leftarrow \text{CrossSourceAttn}(\tilde{\mathbf{h}}_i, \{\tilde{\mathbf{h}}_j \mid j \neq i, m_j = 1\})$ for each valid i
else
  $\hat{\mathbf{h}}_i \leftarrow \tilde{\mathbf{h}}_i$ for each valid i
end if
$\mathbf{z} \leftarrow \text{ResidualProj}(\sum_{i=1}^4 m_i \omega_i \hat{\mathbf{h}}_i)$
return $\mathbf{z}, \mathcal{F}$

Inference proceeds as in Algorithm 1: source representations are projected into their private subspaces, contradiction scores and uncertainties are estimated from the pre-projection representations, and cross-source attention is activated only if the maximum contradiction score exceeds τ.

D Definition of Attention Difference (Δ)

The attention-difference metric in Figure 3 measures the imbalance between intra-modal and inter-modal attention. Large values indicate that a modality is over-focused on itself and under-integrating contradictory information from the other modality, which is characteristic of evidence-averaging failures.

Let $\bar{s}_l$ be mean self-attention and $\bar{c}_l$ be mean cross-attention for a modality at layer l. We compute

$$\Delta = \frac{100\%}{L} \sum_{l=1}^{L} \frac{|\bar{s}_l - \bar{c}_l|}{\bar{s}_l + \epsilon}.$$

A smaller Δ indicates more balanced cross-modal interaction.

E Monte Carlo Dropout Configuration

MACR uses Monte Carlo (MC) dropout to estimate the source-level uncertainty proxy $\mathcal{U}_i$ defined in Eq. (6). During uncertainty estimation, dropout remains active in the lightweight projection heads and the same source representation is evaluated for K stochastic forward passes. The learned model parameters remain fixed, while the dropout masks vary across passes.

For source i, the resulting projected representations $\{\tilde{\mathbf{h}}_i^{(k)}\}_{k=1}^K$ are used to compute their empirical dispersion, which serves as the uncertainty proxy $\mathcal{U}_i$. We use $K = 20$ throughout the experiments. MC dropout is applied only to the projection heads rather than to the full backbone, keeping the additional inference cost limited.

In multi-source settings, $\mathcal{U}_i$ enters the modulation logits in Eq. (8) and can therefore affect the relative allocation of source weights. In the single-source case, the normalized source weight reduces to one, so uncertainty does not alter source allocation.

F Ethical Considerations and Limitations

Ethical Considerations This work studies human-centered affective signals, including physiological and personality-related data in ASCERTAIN. Such technologies carry substantial privacy, bias, and misuse risks, especially in surveillance or high-stakes decision settings. The models and analyses in this paper are intended strictly for research use.

Limitations and Future Work The current evaluation uses existing benchmarks that expose varying degrees of cross-modal disagreement. A natural next step is to construct controlled contradiction benchmarks with explicit conflict/non-conflict splits and counterfactual swap tests, which would allow finer separation of perception failures from arbitration failures. We view this as a promising direction that the present results motivate: the consistent concentration of gains on contradiction-heavy settings suggests that dedicated contradiction-focused evaluation would yield further diagnostic value.

G Efficiency Details and Step-by-Step Case Trace

Training cost and memory. Table 7 reports the computational profile of MACR and of the matched no-attention control (same backbone, data, budget). All numbers are measured on 8×A100-80GB with mixed precision; the extra memory of MACR relative to the control corresponds to the additional projection heads and the MC-dropout forward passes, both of which operate on small heads (∼38M params each) rather than on the backbone.

Table 7: Computational profile on HFM. Latency is per-sample at batch size 1, FP16, single A100. Training time is total wall-clock for one run on 8×A100.

Model	Train (hr)	Peak mem. (GB)	Latency (ms)
MACR, no-attention control	11.2	54.0	23.0
MACR (Full)	**12.4**	**62.3**	**27.0**

MC-dropout forward passes account for less than 15% of total latency ($\leq$20 passes $\times$ 0.2 ms).

Step-by-step trace of a contradiction-heavy sample. To make the recognize–score–arbitrate pipeline fully inspectable, Table 8 reports every intermediate quantity on a representative HFM sarcasm case: the text reads "Oh, that's just great" (literal positive wording used sarcastically) paired with an image of a scowl. This is deliberately one of the easier contradiction cases, so that the trace is not dominated by extreme values; harder cases show the same qualitative pattern with a larger score gap. All quantities are the actual forward-pass values with the trained model at inference time.

Table 8: Numerical trace of a sarcasm case. Step 1: source-specific affect prediction before fusion. Step 2: contradiction and uncertainty estimate. Step 3: arbitration weights $\omega_i \propto \exp(\log \pi_i + s_{\text{base},i} + s_{\text{mod},i})$, before masking/renormalization (somatic and contextual slots are inactive in HFM: $\mathbf{m} = (0, 1, 1, 0)$). Step 4: final prediction. The holistic baseline is the same model with source separation and arbitration removed.

Step	Text (semantic)	Image (visual)	Notes
(1) Source polarity $p(\text{pos}), p(\text{neg})$	0.62 / 0.21	0.19 / 0.58	opposite argmax
(2) Contradiction score $\mathcal{C}_{i,j}$	0.71		$> \tau$=0.3: trigger
(2) Uncertainty $\mathcal{U}_i$	0.18	0.26	visual noisier
(3) $\log \pi_i$	-1.20	-1.20	$\pi = (0.2, 0.3, 0.3, 0.2)$; inactive slots masked
(3) Effective weight ω_i	**0.68**	**0.32**	after masking/renormalization
(4) Predicted affect	positive (sarcastic intent)		Holistic: neutral averaged → wrong

H Additional Results and Representation Geometry

Single-modal emotion benchmarks. Table 9 serves as a robustness check rather than the centerpiece of the paper. It is not intended as a head-to-head fairness claim across identical training regimes, since the grouped baselines include both training-free and trained systems. On single-image emotion

datasets, pairwise cross-source disagreement and arbitration are disabled. These results therefore mainly reflect the behavior of MACR's single-source projected prediction path and are included as a complementary robustness check.

Table 9: Performance on single-modal emotion datasets (vision-only). We report per-dataset accuracy rather than a mixed aggregate.

Method	Backbone	Emotion6	EmoSet	WebEmo7	WebEmo25	Abstract
Training-free baselines (LLaVA-7B)						
Zero-shot	LLaVA-7B	48.32	52.77	25.56	15.71	27.86
Zero-shot-CoT	LLaVA-7B	48.65	51.67	22.93	15.52	26.43
SparseVLM	LLaVA-7B	49.83	54.00	26.64	15.68	26.79
SEPM (β=0.2)	LLaVA-7B	<u>54.04</u>	<u>56.24</u>	<u>42.75</u>	<u>18.34</u>	28.21
Trained specialist (LLaVA-7B)						
MACR (Ours)	LLaVA-7B	**58.47**	**60.31**	**48.52**	**21.78**	**32.15**
Training-free baselines (VILA-8B)						
Zero-shot	VILA-8B	47.47	44.45	41.90	15.35	20.00
SEPM (β=0.1)	VILA-8B	<u>51.35</u>	<u>53.11</u>	<u>44.30</u>	17.36	<u>25.71</u>
Trained specialist (VILA-8B)						
MACR (Ours)	VILA-8B	**56.89**	**57.64**	**47.13**	**22.41**	**29.87**

Extended setting: richer evidence sources (ASCERTAIN). ASCERTAIN exposes somatic and contextual cues absent from the contradiction-heavy sentiment benchmarks. We report it as a transfer test for the recognize-then-arbitrate design, not as primary evidence for the contradiction-specific claim.

Table 10: Performance on ASCERTAIN. We separate affect prediction (arousal/valence) from personality recognition for readability. Averages are computed only within each block.

Method	Arousal	Valence	A/V Avg	Ex	Ag	Con	ES	Op	Pers Avg
Reference baselines from (Subramanian et al., 2018)									
Arousal (Unimodal Avg)	0.588	—	0.588	—	—	—	—	—	—
Valence (Unimodal Avg)	—	0.610	0.610	—	—	—	—	—	—
Personality (Unimodal)	—	—	—	0.563	0.532	0.601	0.730	0.528	0.591
Peripheral (ECG+GSR)	0.620	0.640	0.630	—	—	—	—	—	—
Weighted Fusion (All)	0.640	0.690	0.665	0.650	0.580	0.650	0.770	0.560	0.642
Holistic (Flat Fusion)*	0.667	0.712	0.690	0.673	0.597	0.672	0.782	0.571	0.659
Our method									
MACR (Full)	**0.727**	**0.781**	**0.754**	**0.823**	**0.755**	**0.831**	**0.825**	**0.826**	**0.812**

*Uses MACR features but replaces contradiction-aware arbitration with flat holistic fusion.

Arbitration versus evidence averaging. Table 11 compares MACR with holistic alternatives on multimodal benchmarks with multiple active sources. The gap is especially informative on HFM and TumEmo, where disagreement between text and visual evidence is common.

Representation geometry. Figure 6 provides a representation-level view of the same phenomenon: the baseline embedding geometry is dominated by class overlap, especially around neutral and sarcasm samples, which is exactly where contradiction-sensitive judgments become fragile. Our model organizes the same examples into cleaner local neighborhoods and more separated affect clusters. Positive and negative examples occupy clearer regions, while sarcasm forms a more coherent bridge-like zone instead of being diffusely absorbed by other classes. t-SNE is a qualitative projection, so to avoid resting on visual impressions we quantify the same geometry with a silhouette score and a 10-NN label-consistency measure on the held-out representations (Table 12). MACR (Full) attains a substantially higher silhouette (0.412 vs. 0.288) and a higher 10-NN consistency (67.8% vs. 56.4%), which is consistent with the figure; Fig. 7 supplies the mechanism-level view: panel (a) shows that the learned source manifolds remain well separated after projection, panel (b) shows the desired temporal behavior of the contradiction scorer (accumulating sharply for sarcastic content,

Table 11: Arbitration versus evidence averaging on multimodal benchmarks. The comparison focuses on settings with multiple active sources, where cross-source disagreement and arbitration are well-defined. Improvement is reported relative to MACR (Holistic).

Model Type	Architecture	MVSA-S (F1)	TumEmo (F1)	HFM (F1)
Evidence averaging				
MODA-8B	Modular Duplex Attn	0.705	0.548	0.563
Cambrian-1-8B	Vision Aggregator	0.661	0.344	0.487
MACR (Holistic)*	Flat fusion	0.763	0.697	0.831
Evidence arbitration				
MACR (Full)	Recognize-then-Arbitrate	**0.817**	**0.789±0.005**	**0.892±0.004**
Improvement		**+7.1%**	**+13.2%**	**+7.3%**

*Uses MACR features but removes contradiction-aware arbitration and source separation.

flat for sincere content), and panel (c) shows arbitration weights shifting from a nearly static mixture toward more selective redistribution as conflict intensity increases.

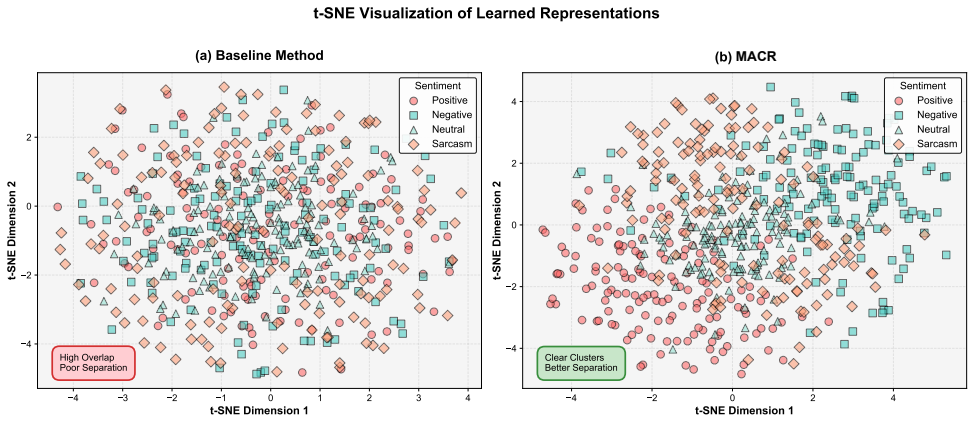


Figure 6: t-SNE visualization of learned affect representations. The baseline produces heavy overlap among auxiliary affect categories derived from HFM samples, indicating that contradiction-sensitive cues are largely entangled in a shared embedding space. By contrast, our model organizes the same examples into substantially cleaner local neighborhoods and more separated affect clusters, which is consistent with the claim that source-preserving arbitration yields a more discriminative geometry for downstream affect understanding.

Table 12: Quantitative representation geometry on HFM held-out test representations. Silhouette and 10-NN consistency are computed in the fused representation space (before the task head).

Model	Silhouette ↑	10-NN same-label ↑	Class-center dist.
MACR (Holistic)	0.288	56.4%	1.13
MACR (Full)	**0.412**	**67.8%**	**1.42**

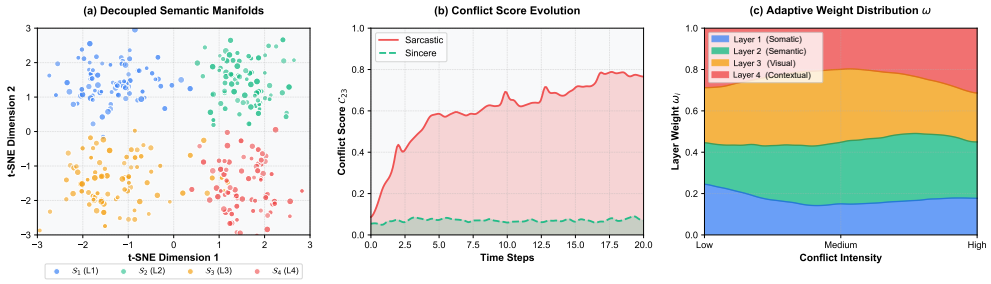


Figure 7: Mechanistic visualization of contradiction-aware reasoning. (a) Source-specific subspaces remain decoupled after projection, avoiding premature mixing of affect evidence. (b) The contradiction score rises steadily for a sarcastic input but stays near zero for a sincere one, indicating that the scorer reacts to genuine cross-source disagreement rather than to sequence length alone. (c) As conflict intensity increases, the arbitration weights shift away from a nearly static mixture toward a more selective redistribution across evidence sources, illustrating that MACR adapts its trust pattern when disagreement becomes salient.

Error decomposition on the HFM test set. Table 13 reports the causal attribution of all 51 errors of the holistic baseline on the full HFM test set under the same annotation protocol as the 480-sample diagnostic split.

Table 13: Error decomposition on the HFM test set (n=494, holistic baseline). Attribution is done per error under the same annotation protocol as the diagnostic split; percentages are computed over all 51 errors.

Attributed cause of error	**# Errors**	**Share**
Cross-modal affective conflict	26	51.0%
Weak / missing perceptual cue	15	29.4%
Label ambiguity / other	10	19.6%
Total baseline errors	**51**	**10.3%** (of 494)

Counterfactual swap test. Table 14 reports the intervention results in detail.

Table 14: Controlled counterfactual swap test on HFM (N=100 paired counterfactuals). The image of each conflict-prone test sample is replaced by an image of opposite polar affect; error rates are reported on the swapped inputs, flip rates as the fraction of samples whose prediction changes under the swap.

Model	**Err. orig.**	**Err. swap**	**Flip rate**	**Δ Err.**
MACR (Holistic)	13.0%	17.8%	21.3%	+4.8 pp
MACR (Full)	**10.3%**	**13.1%**	**15.4%**	+2.8 pp