CARE-KAN: CONFLICT-AWARE EVIDENCE ROUTING FOR ROBUST MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sentiment analysis typically assumes that language, acoustic, and visual modalities provide mutually supportive, complementary evidence. In real-world affective computing, however, modalities frequently diverge due to nuanced emotional expressions, ambient noise, or sensory degradation. Existing methods struggle with this divergence: forcing cross-modal alignment suppresses subtle modality-unique nuances, whereas directly fusing raw conflict features introduces detrimental noise, triggering severe *negative transfer* on clean benchmark data. In this paper, we argue that multimodal conflict is an informative epistemic signal regarding sample difficulty and channel reliability, but it is intrinsically unsafe to exploit naively. To resolve this dilemma, we propose CARE-KAN, a conflict-aware evidence routing framework that shifts multimodal sentiment prediction from unconditional feature fusion to calibrated evidence aggregation. CARE-KAN operates through three coordinated stages: (1) *Tripartite Evidence Decomposition*, which factorizes multimodal features into shared consensus, modality-unique complements, and directional conflict vectors; (2) *Reliability-Adaptive Routing*, which assesses sample-conditioned modality reliability from the joint interaction of these evidence streams to dynamically modulate their predictive weights; and (3) *Zero-Gated Residual Fusion*, which incorporates the calibrated evidence via a zero-centered gate, safeguarding clean-set predictive power while selectively utilizing beneficial disagreement cues. The resulting representation is fed into a Kolmogorov–Arnold Network (KAN) prediction head to capture complex non-smooth residual decision surfaces. Evaluations on four benchmarks (CMU-MOSI, CMU-MOSEI, CH-SIMS, and IEMOCAP) show that CARE-KAN eliminates negative transfer on clean data while demonstrating superior resilience under acoustic/visual Gaussian noise and modality dropout. Mechanistic analyses further reveal that conflict magnitude reliably tracks prediction difficulty, and our router dynamically attenuates corrupted channels while preserving salient task-relevant evidence.

1 INTRODUCTION

Multimodal sentiment analysis (MSA) integrates verbal language, vocal prosody, and facial expressions to decipher human affective states (Zadeh et al., 2016; 2018; Liang et al., 2024). Predominant paradigms, spanning cross-modal attention (Tsai et al., 2019), contrastive alignment (Yu et al., 2021), subspace disentanglement (Hazarika et al., 2020; Yang et al., 2022), and knowledge distillation (Li et al., 2023), rely on the core assumption of multimodal complementarity (Baltrušaitis et al., 2018). However, these approaches implicitly presuppose harmonious cross-modal consensus, treating feature alignment as an unconditional objective and presuming all sensory streams to be uniformly reliable across samples.

Motivation: The Dilemma of Multimodal Disagreement. In practice, cross-modal consensus is routinely violated: human affect exhibits layered incongruities like sarcasm (Zadeh et al., 2018; Zhu et al., 2024), while sensor degradation (e.g., acoustic noise or illumination shifts) introduces corruptive discrepancies (Liu et al., 2024a). Confronted with such divergence, existing paradigms oscillate between two flawed extremes. On the one hand, *over-alignment* frameworks (Hazarika et al., 2020; Yan et al., 2026) penalize distribution divergences to enforce a shared space; this indiscriminate homogenization treats divergence as noise, inadvertently erasing subtle, modality-unique

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.

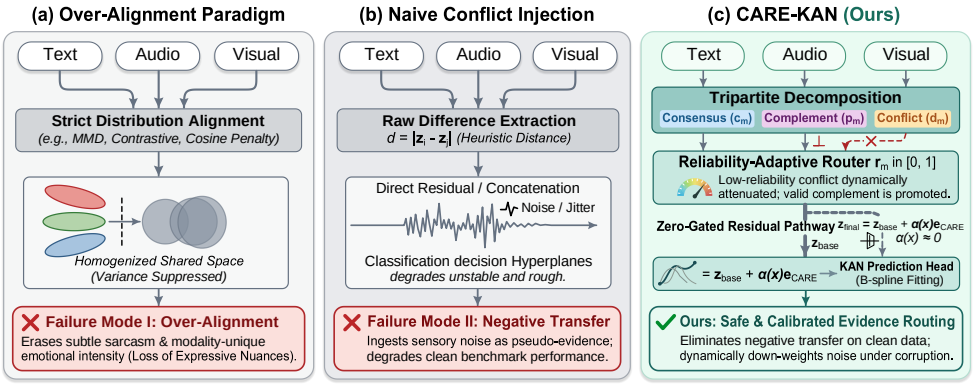

Figure 1: **Conceptual comparison of multimodal disagreement paradigms.** (a) *Over-Alignment* suppresses modality-unique expressive nuances. (b) *Naive Conflict Injection* triggers severe negative transfer on clean data. (c) *CARE-KAN (Ours)* adaptively routes calibrated evidence for safe and robust multimodal prediction.

expressive nuances. On the other hand, *naive residual fusion* directly injects uncalibrated conflict features into downstream regressors, forcing models to ingest sensory noise as pseudo-evidence and triggering severe **negative transfer** on clean benchmark data (Wei et al., 2024). This exposes a central dilemma: *How can a multimodal system distinguish informative affective divergence from corrupted sensory noise, and harness conflict without suffering negative transfer?*

Our Approach: CARE-KAN. We argue that multimodal conflict is an **informative epistemic signal** diagnosing sample difficulty and channel fidelity, but it cannot be exploited unconditionally. Rather than enforcing over-alignment (Figure 1a) or risking negative transfer (Figure 1b), we propose **CARE-KAN**, a **C**onflict-**A**ware **R**eliability-guided **E**vidence routing framework equipped with a **K**olmogorov–**A**rnold **N**etwork (KAN) head (Liu et al., 2024b) (Figure 1c). CARE-KAN establishes a disciplined three-stage evidence aggregation hierarchy: (1) *Tripartite Evidence Decomposition* factorizes unimodal representations beyond coarse shared-private partitions (Zhou et al., 2025b; Wang et al., 2025b) into consensus ($\boldsymbol{c}_m$), orthogonal complement ($\boldsymbol{p}_m$), and directional conflict ($\boldsymbol{d}_m$) vectors; (2) *Reliability-Adaptive Routing* moves beyond heuristic rebalancing (Guo et al., 2025; Wei et al., 2025) to assess sample-level channel reliability ($r_m \in [0, 1]$) from the joint interaction of all three streams, dynamically attenuating corrupted channels while preserving reliable cues; and (3) *Zero-Gated Residual Fusion* channels calibrated residuals through a zero-initialized gate ($\alpha(\boldsymbol{x}) \approx 0$), which provably shields clean-set predictions from residual noise at initialization (Section 3.4) while selectively mobilizing conflict evidence under sensory ambiguity. The representation is then mapped via a KAN head whose localized B-splines fit non-smooth residual decision surfaces.

Evaluations on four benchmarks show that CARE-KAN mitigates clean-set negative transfer while achieving substantially smaller performance degradation (ΔMAE, Table 2) under Gaussian noise and frame dropout. Diagnostic analyses further verify that conflict magnitude reliably tracks prediction difficulty, and that our router dynamically suppresses corrupted channels.

The primary contributions of this work are summarized as follows:

- **Conceptual Paradigm Shift**: We reframe multimodal conflict from an unwanted nuisance into an informative epistemic signal, formally exposing the negative transfer vulnerability caused by naive divergence usage.
- **Disciplined Routing Architecture**: We propose CARE-KAN, integrating tripartite decomposition, reliability-adaptive routing, and zero-gated residual fusion with a localized B-spline KAN prediction head to exploit conflict safely.
- **Empirical Validation Across Four Benchmarks**: On four benchmarks, CARE-KAN matches or exceeds the strongest baseline on clean data (Table 1) while more than halving corruption-induced degradation relative to DecAlign (Table 2), supported by diagnostic analyses of conflict, reliability, and ablation (Table 3, Figure 3).

2 Related Work

2.1 Multimodal Affective Analysis and Representation Disentanglement

Early multimodal sentiment analysis relied on directional cross-attention (Tsai et al., 2019) or multi-task contrastive learning (Yu et al., 2021), where uncoordinated fusion often allowed dominant modalities to suppress weaker channels. To alleviate feature entanglement, representation disentanglement established the classical "shared–private" factorization scheme via distribution disparity and reconstruction losses (Hazarika et al., 2020; Yang et al., 2022). Recent studies have extended disentanglement to fine-grained contexts (Sun et al., 2024; Zhang et al., 2024a) and translation-based autoencoding (Zeng et al., 2024). The 2025–2026 literature further refines these boundaries via language-focused attractors (Wang et al., 2025b), geometric bounds (Zhou et al., 2025b), representation shifting (Zhao et al., 2025), controlled information criteria (Wang et al., 2025a), and mixture-of-disentangled-experts frameworks (Li et al., 2026; Chang et al., 2026). However, existing frameworks strictly adhere to a binary categorical assumption, presuming multimodal features are either mutually shared or strictly private, which breaks down when cross-modal affective conflict occurs. To overcome this limitation, our work introduces *Tripartite Evidence Decomposition*, explicitly isolating directional conflict as an independent geometric component alongside consensus and orthogonal complements to prevent manifold pollution.

2.2 Multimodal Disagreement, Robustness, and Negative Transfer

To withstand sensory imperfections such as acoustic noise and channel dropouts (Liu et al., 2024a; Dong et al., 2024), robust multimodal learning has explored cross-modal translation (Liu et al., 2024a; Zeng et al., 2024), self-distillation (Li et al., 2023; 2024b), and dynamic fusion (Zhang et al., 2024b). Concurrently, studies in 2024–2025 have revealed that uncoordinated multimodal optimization suffers from severe gradient conflicts and *negative transfer* (Wei et al., 2024; Hua et al., 2024), prompting regularization strategies such as on-the-fly gradient modulation (Wei et al., 2024), classifier-guided re-weighting (Guo et al., 2025), diagnostic re-learning (Wei et al., 2025), and adaptive unimodal regulation (Huang et al., 2025). Nevertheless, existing paradigms predominantly treat cross-modal divergence as an undesirable pathology to be suppressed, failing to recognize that affective disagreement frequently conveys intentional communicative nuances (e.g., sarcasm), while naive difference injection triggers severe negative transfer. In contrast, CARE-KAN reframes conflict as an epistemic diagnostic asset, deploying a *Reliability-Adaptive Router* and *Zero-Gated Residual Fusion* to safely harvest beneficial disagreement without compromising clean-set fidelity.

2.3 Kolmogorov–Arnold Networks in Representation Learning

Kolmogorov–Arnold Networks (KANs) (Liu et al., 2024b) provide an expressive alternative to MLPs by deploying learnable univariate functions (parameterized via B-splines) directly on network edges, achieving high parameter efficiency while avoiding catastrophic forgetting. Recent adaptations span spatial vision modeling (Li et al., 2024a), dynamic time-series forecasting (Xu et al., 2024; Zhou et al., 2025a), and multimodal affective computing (Luo et al., 2025; Shen & Ma, 2026), showing that KAN can capture high-order nonlinear inter-modal correlations. While existing multimodal applications predominantly treat KAN as a generic plug-and-play substitute, standard MLPs exhibit spectral bias that over-smooths sparse, sample-dependent conflict residuals $\boldsymbol{e}_{\text{CARE}}$. CARE-KAN exploits the *compact local support* of KAN's B-spline basis functions, utilizing localized function approximation to adaptively fit sharp, irregular residual decision boundaries without disturbing the smooth primary base manifold.

3 Methodology

3.1 Overview and Primary Representation Pipeline

Let $\mathbb{D} = \{(\boldsymbol{X}_n, y_n)\}_{n=1}^{N}$ denote a multimodal affective dataset where $y_n \in \mathbb{R}$ is the ground-truth target and $\boldsymbol{X}_n = \{\boldsymbol{X}_t, \boldsymbol{X}_a, \boldsymbol{X}_v\}$ comprises sequential observations across language (t), acoustic (a), and visual (v) modalities. For each modality $m \in \mathbb{M} = \{t, a, v\}$ with raw features $\boldsymbol{X}_m \in \mathbb{R}^{T_m \times d_m}$, we apply modality-specific 1D temporal convolutions to project inputs into a standardized space of

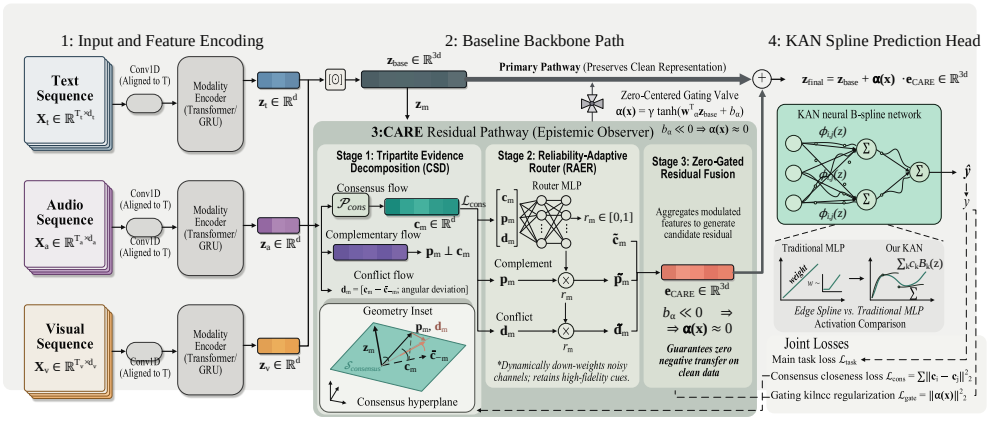

Figure 2: **The overall architectural framework of CARE-KAN.** Unimodal inputs are standardized via temporal convolutions and mapped to bottleneck representations $\boldsymbol{z}_m$, constructing the primary base representation $\boldsymbol{z}_{\text{base}}$. In parallel, the CARE pathway orchestrates: (1) *Tripartite Evidence Decomposition* into consensus ($\boldsymbol{c}_m$), orthogonalized complement ($\boldsymbol{p}_m$), and directional conflict ($\boldsymbol{d}_m$); (2) *Reliability-Adaptive Routing* estimating channel fidelity $r_m \in [0, 1]$ to modulate evidence flow; and (3) *Zero-Gated Residual Fusion* injecting calibrated evidence $\boldsymbol{e}_{\text{CARE}}$ via a zero-centered gate $\alpha(\boldsymbol{x})$. The combined representation $\boldsymbol{z}_{\text{final}}$ is decoded into affective predictions by a localized B-spline KAN prediction head.

sequence length T and hidden dimension d, yielding $\tilde{\boldsymbol{X}}_m \in \mathbb{R}^{T\times d}$. Unimodal temporal encoders followed by temporal pooling compress each channel into an instance-level bottleneck vector $\boldsymbol{z}_m \in \mathbb{R}^d$. Concatenating these features forms the primary base representation:

$$\boldsymbol{z}_{\text{base}} = [\boldsymbol{z}_t^\top, \boldsymbol{z}_a^\top, \boldsymbol{z}_v^\top]^\top \in \mathbb{R}^{3d}. \tag{1}$$

Dual-Pathway Architectural Philosophy. Mainstream fusion paradigms project multimodal inputs into an unconstrained, entangled space (Tsai et al., 2019), inadvertently forcing models to absorb uncalibrated sensory discrepancies and triggering severe negative transfer on clean data (Wei et al., 2024). CARE-KAN circumvents this pathology via an asymmetric dual-pathway routing paradigm (Figure 2). The primary pathway directly propagates $\boldsymbol{z}_{\text{base}}$ to anchor core predictive power on clean inputs. Concurrently, the CARE pathway functions as a sample-conditioned epistemic observer: it isolates cross-modal divergence via *Tripartite Evidence Decomposition* (Section 3.2), assesses channel trustworthiness via *Reliability-Adaptive Routing* (Section 3.3), and synthesizes a calibrated residual evidence vector $\boldsymbol{e}_{\text{CARE}}$ channeled through a zero-centered gate $\alpha(\boldsymbol{x})$ (Section 3.4) to the KAN prediction head (Section 3.5). The composite feature $\boldsymbol{z}_{\text{final}} = \boldsymbol{z}_{\text{base}} + \alpha(\boldsymbol{x}) \cdot \boldsymbol{e}_{\text{CARE}}$ preserves the stability of the pristine base manifold at initialization (Appendix A) while safely mobilizing informative conflict evidence under sensory ambiguity.

3.2 Tripartite Evidence Decomposition

To transcend the expressiveness bottlenecks of coarse shared-private partitions, we explicitly factorize each unimodal vector $\boldsymbol{z}_m$ into consensus ($\boldsymbol{c}_m$), orthogonal complement ($\boldsymbol{p}_m$), and directional conflict ($\boldsymbol{d}_m$). We refer to this three-term decomposition as **C3D** (Consensus–Complement–Conflict Decomposition) in subsequent sections.

Consensus Extraction. A shared nonlinear projection $\mathcal{P}_{\text{com}} : \mathbb{R}^d \to \mathbb{R}^d$ maps $\boldsymbol{z}_m$ to a common semantic manifold: $\boldsymbol{c}_m = \mathcal{P}_{\text{com}}(\boldsymbol{z}_m) \in \mathbb{R}^d$. Mutual affective agreement is enforced via a pairwise consensus alignment objective:

$$\mathcal{L}_{\text{cons}} = \frac{2}{M(M-1)} \sum_{1\le i<j\le M} \|\boldsymbol{c}_i - \boldsymbol{c}_j\|_2^2, \tag{2}$$

where $M = |\mathbb{M}| = 3$. Minimizing $\mathcal{L}_{\text{cons}}$ establishes a stable geometric anchor across disparate modalities.

Orthogonal Complement Extraction. Rather than relying on soft statistical penalties (Hazarika et al., 2020) that suffer from information leakage, we enforce closed-form geometric projection. Specifically, we project $\boldsymbol{z}_m$ onto the orthogonal complement of the consensus subspace $\mathbb{S}_{\text{consensus}} = \text{span}(\boldsymbol{c}_m)$ via the Gram–Schmidt operator:

$$\boldsymbol{p}_m = \boldsymbol{z}_m - \frac{\langle \boldsymbol{z}_m, \boldsymbol{c}_m \rangle}{\|\boldsymbol{c}_m\|_2^2 + \epsilon} \boldsymbol{c}_m \in \mathbb{R}^d, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes Euclidean inner product and $\epsilon > 0$ ensures numerical stability. By construction, $\langle \boldsymbol{p}_m, \boldsymbol{c}_m \rangle \to 0$ as $\epsilon \to 0^+$ (see proof in Appendix A), providing a near-orthogonal geometric projection that suppresses per-sample consensus leakage in the complement representation.

Directional Conflict Formulation. We define the leave-one-out consensus center for modality m as $\bar{\boldsymbol{c}}_{-m} = \frac{1}{M-1} \sum_{j \neq m} \boldsymbol{c}_j$. To avoid discarding spatial orientation via heuristic scalar differences, conflict is characterized by jointly coupling spatial translational displacement with cosine angular divergence:

$$\boldsymbol{d}_m = \boldsymbol{W}_d \left[(\boldsymbol{c}_m - \bar{\boldsymbol{c}}_{-m})^\top,\ 1 - \frac{\boldsymbol{c}_m^\top \bar{\boldsymbol{c}}_{-m}}{\|\boldsymbol{c}_m\|_2 \|\bar{\boldsymbol{c}}_{-m}\|_2 + \epsilon} \right]^\top + \boldsymbol{b}_d \in \mathbb{R}^d, \tag{4}$$

where $\boldsymbol{W}_d \in \mathbb{R}^{d \times (d+1)}$ and $\boldsymbol{b}_d \in \mathbb{R}^d$ are learnable projection weights. Thus, Tripartite Decomposition yields an expressive evidence triplet $\mathcal{E}_m = \{\boldsymbol{c}_m, \boldsymbol{p}_m, \boldsymbol{d}_m\}$ for each modality, establishing a structured foundation for reliability calibration.

3.3 Reliability-Adaptive Evidence Routing

Unconditional fusion implicitly presumes all sensory channels remain uniformly dependable across all instances, leaving models acutely vulnerable to sporadic sensory degradation. We refer to this reliability-estimation and evidence-modulation mechanism as **RAER** (Reliability-Adaptive Evidence Routing) in subsequent sections.

Sample-Conditioned Reliability Estimation. To adjudicate whether an observed divergence represents authentic affective nuance (e.g., sarcasm) or uninformative sensor corruption, our router evaluates a continuous fidelity score $r_m \in (0, 1)$ conditioned on the joint interaction of all three evidence streams $\boldsymbol{e}_m = [\boldsymbol{c}_m^\top, \boldsymbol{p}_m^\top, \boldsymbol{d}_m^\top]^\top \in \mathbb{R}^{3d}$:

$$r_m = \sigma \left(\frac{\boldsymbol{w}_u^\top \text{LeakyReLU}(\boldsymbol{W}_e \boldsymbol{e}_m + \boldsymbol{b}_e) + b_u}{\tau} \right), \tag{5}$$

where $\boldsymbol{W}_e \in \mathbb{R}^{d \times 3d}$, $\boldsymbol{b}_e, \boldsymbol{w}_u \in \mathbb{R}^d$, $b_u \in \mathbb{R}$, and $\tau > 0$ is a temperature hyperparameter controlling routing sharpness (sensitivity analysis in Appendix D.1). Crucially, authentic affective nuances exhibit coherent semantic structure in $\boldsymbol{p}_m$ alongside structured displacement $\boldsymbol{d}_m$; conversely, hardware-induced corruption triggers high-entropy fluctuations in $\boldsymbol{p}_m$ paired with uncoordinated, excessive divergence $\boldsymbol{d}_m$, naturally driving $r_m \to 0$.

Evidence Modulation and Residual Assembly. The router dynamically modulates evidence flows based on r_m:

$$\tilde{\boldsymbol{p}}_m = r_m \cdot \boldsymbol{p}_m, \qquad \tilde{\boldsymbol{d}}_m = r_m \cdot \boldsymbol{d}_m, \qquad \tilde{\boldsymbol{c}}_m = \frac{\exp(r_m/\tau)}{\sum_j \exp(r_j/\tau)} \cdot \boldsymbol{c}_m. \tag{6}$$

Eq. (6) amplifies dependable channel-unique cues while actively attenuating noise-induced conflicts. Modulated features are concatenated into $\tilde{\mathcal{E}} \in \mathbb{R}^{9d}$ and projected to form the candidate residual evidence:

$$\boldsymbol{e}_{\text{CARE}} = \text{LayerNorm} \left(\boldsymbol{W}_r \tilde{\mathcal{E}} + \boldsymbol{b}_r \right) \in \mathbb{R}^{3d}, \tag{7}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{3d \times 9d}$ and $\boldsymbol{b}_r \in \mathbb{R}^{3d}$, matching the dimensionality of $\boldsymbol{z}_{\text{base}}$.

3.4 Zero-Gated Residual Fusion for Negative Transfer Mitigation

Directly superimposing unconstrained conflict features ($\boldsymbol{z}_{\text{final}} = \boldsymbol{z}_{\text{base}} + \boldsymbol{e}_{\text{CARE}}$) causes negative transfer on clean data by injecting uncalibrated variance into the decision manifold. To provide an

unconditional theoretical safeguard, we regulate residual injection via a sample-conditioned scalar gate $\alpha(\boldsymbol{x})$:

$$\boldsymbol{z}_{\text{final}} = \boldsymbol{z}_{\text{base}} + \alpha(\boldsymbol{x}) \cdot \boldsymbol{e}_{\text{CARE}} \in \mathbb{R}^{3d}, \qquad \alpha(\boldsymbol{x}) = \gamma \cdot \sigma\left(\boldsymbol{w}_\alpha^\top \boldsymbol{z}_{\text{base}} + b_\alpha\right) \in (0, \gamma), \tag{8}$$

where $\boldsymbol{w}_\alpha \in \mathbb{R}^{3d}$, $b_\alpha \in \mathbb{R}$, and $\gamma \in (0, 1]$. We enforce a negative bias initialization protocol: setting $b_\alpha \leftarrow b_0 = -3.0$ and initializing $\boldsymbol{w}_\alpha \sim \mathcal{N}(0, 10^{-4})$ guarantees $\alpha(\boldsymbol{x}) \approx 0$ at the onset of training, strictly enforcing $\boldsymbol{z}_{\text{final}} \approx \boldsymbol{z}_{\text{base}}$.

Gradient Dynamics and Insulation. The analytical gradient with respect to the base representation expands as $\frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{base}}} = \frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{final}}} \left(\boldsymbol{I} + \frac{\partial \alpha(\boldsymbol{x})}{\partial \boldsymbol{z}_{\text{base}}} \boldsymbol{e}_{\text{CARE}}^\top + \alpha(\boldsymbol{x}) \frac{\partial \boldsymbol{e}_{\text{CARE}}}{\partial \boldsymbol{z}_{\text{base}}} \right)$ (see derivation in Appendix A). Under our saturated negative bias regime ($b_\alpha \ll 0$), both $\alpha(\boldsymbol{x}) \to 0$ and $\sigma'(b_\alpha) \to 0$, causing cross-terms to vanish completely:

$$\lim_{b_\alpha \to -\infty} \frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{base}}} = \frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{final}}} \cdot \boldsymbol{I}. \tag{9}$$

Eq. (9) proves that at initialization, the task-loss gradient flowing back through the residual branch vanishes, protecting the base backbone from uncalibrated residual interference during early training. The gate is driven open via $\frac{\partial \mathcal{L}_{\text{task}}}{\partial \alpha(\boldsymbol{x})} = (\frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{final}}})^\top \boldsymbol{e}_{\text{CARE}}$ only when residual evidence provides true empirical risk reduction. To enforce parsimony on clean samples, we incorporate an explicit shrinkage regularization:

$$\mathcal{L}_{\text{gate}} = \frac{1}{|\mathbb{B}|} \sum_{\boldsymbol{x} \in \mathbb{B}} \|\alpha(\boldsymbol{x})\|_2^2. \tag{10}$$

3.5 Kolmogorov–Arnold Network Prediction and Joint Optimization

To map $\boldsymbol{z}_{\text{final}}$ to target outputs without inducing over-smoothing, we instantiate the prediction head via a Kolmogorov–Arnold Network (KAN) (Liu et al., 2024b). Grounded in the Kolmogorov–Arnold representation theorem, KAN replaces static node activations with learnable univariate functions on network edges. Let $\boldsymbol{z}^{(0)} = \boldsymbol{z}_{\text{final}} \in \mathbb{R}^{d_0}$ ($d_0 = 3d$). A KAN layer updates representations via learnable univariate B-splines:

$$\boldsymbol{z}_i^{(l)} = \sum_{j=1}^{d_{l-1}} \phi_{l,i,j}\left(\boldsymbol{z}_j^{(l-1)}\right), \quad \phi_{l,i,j}(x) = w_{b,l,i,j} \cdot \text{SiLU}(x) + w_{s,l,i,j} \sum_{n=1}^{G+k_o} c_{l,i,j,n} B_n(x), \tag{11}$$

where $\text{SiLU}(x)$ supplies smooth base propagation, $B_k(x)$ denotes standard Cox–de Boor B-splines of order k_o over G grid intervals, $c_k \in \mathbb{R}$ are trainable coefficients, and w_b, w_s are scaling factors. Stacking L layers yields the prediction $\hat{y} = \text{KAN}(\boldsymbol{z}_{\text{final}})$.

Theoretical Synergy Between CARE and KAN. Standard MLPs rely on global activations with spectral bias, which tend to over-smooth localized variations and wash out sparse, sample-dependent conflict residuals $\boldsymbol{e}_{\text{CARE}}$. In contrast, B-spline basis functions possess *compact local support*: $B_k(x) = 0$ for all $x \notin [t_k, t_{k+k_o+1}]$, where $\{t_k\}$ are grid knot points. This locality mathematically ensures that adjusting spline parameters alters the function landscape exclusively within a bounded neighborhood. Consequently, KAN flexibly accommodates the sharp, localized residual adjustments delivered by $\alpha(\boldsymbol{x})\boldsymbol{e}_{\text{CARE}}$ on ambiguous instances, while preserving the unperturbed global manifold of $\boldsymbol{z}_{\text{base}}$ on clean samples.

End-to-End Joint Objective. The entire framework is trained jointly via:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}}, \tag{12}$$

where $\lambda_{\text{cons}}, \lambda_{\text{gate}} > 0$ balance task precision, consensus consistency (Eq. 2), and gate parsimony (Eq. 10). The task-specific loss $\mathcal{L}_{\text{task}}$ is formulated as:

$$\mathcal{L}_{\text{task}} = \begin{cases} \dfrac{1}{|\mathbb{B}|} \displaystyle\sum_{i \in \mathbb{B}} (y_i - \hat{y}_i)^2, & \text{(Continuous Sentiment Regression)} \\ -\dfrac{1}{|\mathbb{B}|} \displaystyle\sum_{i \in \mathbb{B}} \sum_{c=1}^{C_0} y_{i,c} \log \hat{p}_{i,c}, & \text{(Categorical Emotion Recognition)} \end{cases} \tag{13}$$

where $C_0 = 6$ classes for IEMOCAP, and $\mathbb{B}$ denotes the mini-batch. Eq. (12) ensures joint optimization of accuracy, consensus alignment, and gate parsimony, enabling CARE-KAN to maintain state-of-the-art predictive fidelity while establishing robust adaptation under sensory divergence.

Table 1: **Performance comparison across four multimodal benchmarks on clean data.** We report core metrics for all four datasets, with full five-metric versions in Appendix C. Best results are in **bold**; second-best are underlined.

Method	CMU-MOSI			CMU-MOSEI			IEMOCAP (six-class)		CH-SIMS	
	MAE ↓	Acc-2 ↑	F1 ↑	MAE ↓	Acc-2 ↑	F1 ↑	WAcc ↑	WAF1 ↑	MAE ↓	F1 ↑
MFM (Tsai et al., 2018)	0.951	78.18	78.10	0.681	78.93	76.45	63.38	63.41	0.471	75.28
MulT (Tsai et al., 2019)	0.846	81.70	81.66	0.673	80.85	80.86	65.53	65.21	0.455	76.96
MISA (Hazarika et al., 2020)	0.788	82.07	82.43	0.594	82.03	82.13	68.48	68.25	0.437	78.43
Self-MM (Yu et al., 2021)	0.765	82.88	83.04	0.576	82.43	82.47	70.35	70.43	0.432	77.97
FDMER (Yang et al., 2022)	0.760	83.01	83.22	0.571	83.88	83.35	71.33	71.17	0.424	78.74
DMD (Li et al., 2023)	0.744	83.24	83.55	0.561	84.17	83.88	72.03	71.98	0.421	79.88
ReconBoost (Hua et al., 2024)	0.793	82.59	82.72	0.599	82.98	83.14	71.44	71.58	0.413	80.41
CGGM (Guo et al., 2025)	0.787	82.73	82.89	0.584	83.72	83.94	72.25	72.17	0.417	80.12
DecAlign (Qian et al., 2026)	0.735	85.75	85.82	**0.543**	86.48	86.07	73.35	73.43	0.403	81.85
Base Backbone	0.738	83.85	83.92	0.558	84.62	84.55	72.50	72.40	0.415	80.60
Plain-Residual KAN	0.782	81.15	81.20	0.605	82.10	82.35	70.15	70.20	0.442	78.30
CARE-KAN (Ours)	**0.687**	**87.02**	**87.06**	0.553	**87.04**	**86.99**	**74.12**	**74.20**	**0.391**	**82.45**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. We benchmark CARE-KAN on four standard affective computing datasets: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), and IEMOCAP (Busso et al., 2008); full dataset statistics, splits, and label ranges are provided in Appendix B.1 (Table B.1). For regression (MOSI, MOSEI, CH-SIMS), we evaluate Mean Absolute Error (MAE ↓) and Pearson Correlation (Corr ↑). For classification, we report binary accuracy (Acc-2 ↑), binary F1-score ↑, and 7-class accuracy (Acc-7 ↑). On IEMOCAP, we evaluate Weighted Accuracy (WAcc ↑) and Weighted F1 (WAF1 ↑) to address class imbalance.

Baselines and Protocols. We benchmark against 9 representative methods across four families: early/attention fusion (MFM (Tsai et al., 2018), MulT (Tsai et al., 2019)), disentanglement (MISA (Hazarika et al., 2020), FDMER (Yang et al., 2022), DecAlign (Qian et al., 2026)), contrastive/distillation (Self-MM (Yu et al., 2021), DMD (Li et al., 2023)), and modality balancing (ReconBoost (Hua et al., 2024), CGGM (Guo et al., 2025)), alongside two internal controls: *Base Backbone*, which maps $\boldsymbol{z}_{\text{base}}$ (Eq. 1) directly through the KAN head without the CARE pathway, and *Plain-Residual KAN*, which fixes $\alpha(\boldsymbol{x}) \equiv 1$ in Eq. 8 to inject the unmodulated residual $\boldsymbol{e}_{\text{CARE}}$ without RAER or zero-gating. Features are extracted using MMSA-FET (BERT, COVAREP, OpenFace) and DeBERTa/wav2vec for IEMOCAP, standardizing sequences to $T = 50, d = 128$ via 1D convolutions. The KAN head uses a 2-layer architecture (one hidden layer, dimensions $[384, 128, 1]$) with cubic B-splines ($k_o = 3, G = 5$). Gating parameters are $\gamma = 0.1, b_0 = -3.0$, with loss weights $\lambda_{\text{cons}} = 0.05, \lambda_{\text{gate}} = 0.01$. Models are optimized via AdamW with cosine annealing (learning rate 5×10^{-5} to 1×10^{-4}) for 50 epochs over five random seeds.

4.2 QUANTITATIVE RESULTS ON CLEAN BENCHMARKS

Following Qian et al. (2026), Table 1 reports core metrics across all four datasets, with full five-metric evaluations (including Corr and Acc-7) documented in Appendix C. CARE-KAN sets new state-of-the-art performance across benchmarks.

Benchmark Superiority and Generalization. On CMU-MOSI, CARE-KAN establishes new state-of-the-art benchmarks, reducing MAE to **0.6868** (an absolute drop of 0.0482 over DecAlign's 0.735) while elevating Acc-2 and F1 to **87.02%** ($+1.27\%$) and **87.06%** ($+1.24\%$), alongside superior correlation (Corr = **0.8333**, Acc-7 = 44.67%). On CMU-MOSEI, CARE-KAN achieves leading classification accuracy (Acc-2 = **87.04%**, F1 = **86.99%**) and correlation (Corr = **0.7722**, Acc-7 = 52.23%) with competitive regression (MAE = 0.5527). These gains extend to bilingual sentiment and categorical emotion tasks: CARE-KAN sets leading results on Chinese CH-SIMS (MAE = **0.3912**, F1 = **82.45%**) and six-class IEMOCAP (WAcc = **74.12%**, WAF1 = **74.20%**), indicating that reliability-adaptive routing generalizes across distinct linguistic structures and affective granularities.

Empirical Diagnosis of Negative Transfer. Comparing our internal controls in Table 1 reveals the role of zero-gating. Directly injecting uncalibrated conflict features (*Plain-Residual KAN*) degrades

Table 2: **Robustness stress-testing under sensory corruptions on CMU-MOSI.** We report performance degradation $\Delta\text{MAE} = \text{MAE}_{\text{corrupt}} - \text{MAE}_{\text{clean}}$ across escalating visual noise (σ) and acoustic frame dropout (p). Best results are in **bold**; second-best are underlined.

Method	Visual Gaussian Noise (ΔMAE ↓)			Acoustic Frame Dropout (ΔMAE ↓)		
	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 1.0$	$p = 0.3$	$p = 0.5$	$p = 0.7$
MulT (Tsai et al., 2019)	+0.078	+0.162	+0.285	+0.065	+0.138	+0.224
MISA (Hazarika et al., 2020)	+0.062	+0.134	+0.245	+0.051	+0.112	+0.188
DMD (Li et al., 2023)	+0.048	+0.105	+0.198	+0.043	+0.095	+0.162
CGGM (Guo et al., 2025)	+0.052	+0.112	+0.209	+0.045	+0.098	+0.168
DecAlign (Qian et al., 2026)	<u>+0.042</u>	<u>+0.092</u>	<u>+0.176</u>	<u>+0.038</u>	<u>+0.084</u>	<u>+0.149</u>
Base Backbone	+0.045	+0.098	+0.182	+0.039	+0.089	+0.151
Plain-Residual KAN	+0.084	+0.178	+0.312	+0.072	+0.155	+0.258
CARE-KAN (Ours)	**+0.018**	**+0.041**	**+0.076**	**+0.015**	**+0.035**	**+0.068**

performance across all datasets: on CMU-MOSI, MAE worsens from 0.738 to 0.782 and Acc-2 drops by 2.70%; on CMU-MOSEI, MAE rises to 0.605. This pattern indicates that unconstrained conflict features inject optimization noise, causing negative transfer on clean data. CARE-KAN avoids this failure mode: rather than degrading relative to the Base Backbone, it improves over it, turning conflict into an informative asset that elevates MOSI Acc-2 by $+3.17\%$ over the base backbone.

4.3 Robustness Stress-Testing under Sensory Corruption

We evaluate model resilience on CMU-MOSI under two corruption protocols during inference: (1) *Additive Gaussian Noise* on visual features with variance $\sigma \in \{0.2, 0.5, 1.0\}$; and (2) *Temporal Frame Dropout* on acoustic sequences with dropout probability $p \in \{0.3, 0.5, 0.7\}$. We quantify degradation using $\Delta\text{MAE} = \text{MAE}_{\text{corrupt}} - \text{MAE}_{\text{clean}}$, where lower indicates superior resilience.

Resilience Analysis and Circuit-Breaker Principle. As summarized in Table 2, conventional architectures show severe performance collapse under corruption: MulT deteriorates by $+0.285$ at $\sigma = 1.0$, consistent with unfiltered noise propagating through its attention-based fusion, while DecAlign ($+0.176$) and MISA ($+0.245$) show comparable degradation, plausibly because their consensus-alignment objectives force noisy inputs into the shared space. Plain-Residual KAN is the most vulnerable ($+0.312$), as its unguided residual pathway amplifies sensory noise into the decision boundary. CARE-KAN bounds degradation to **+0.041** at $\sigma = 0.5$ and **+0.076** at $\sigma = 1.0$ (and **+0.068** at $p = 0.7$), outperforming DecAlign by more than two-fold. This pattern is consistent with an *adaptive circuit-breaker principle*: when a channel degrades, the router detects high-entropy complement $\boldsymbol{p}_m$ and erratic conflict $\boldsymbol{d}_m$, driving $r_m \to 0$ to attenuate compromised residuals ($\tilde{\boldsymbol{d}}_m, \tilde{\boldsymbol{p}}_m \to \mathbf{0}$) while zero-gating shields the unperturbed base manifold.

4.4 Diagnostic and Mechanistic Analysis

We conduct diagnostic analyses verifying the relationship between conflict and difficulty, as well as dynamic router behavior under corruption.

Conflict Tracks Sample Prediction Difficulty. Stratifying the CMU-MOSI test set into equal-sized tertiles based on aggregate conflict norm $\|\boldsymbol{d}\|_2 = \frac{1}{M}\sum_m \|\boldsymbol{d}_m\|_2$ reveals that regression error increases monotonically with conflict magnitude across all models (Figure 3a). Base Backbone and DecAlign error rates climb from ~ 0.62 in the Low Conflict tertile to > 0.86 in the High Conflict tertile, consistent with cross-modal conflict tracking affective ambiguity (e.g., sarcasm). While baselines escalate sharply, CARE-KAN restricts high-conflict MAE to **0.758** (a 0.110 reduction over DecAlign), indicating that selective routing through localized KAN splines converts difficult conflict into discriminative cues.

Dynamic Channel Attenuation and Case Study. Tracking internal routing activations under escalating visual noise ($\sigma \in [0, 1.0]$) in Figure 3b shows that the visual reliability score r_v plunges from 0.88 to 0.05, dynamically cutting off corrupted visual residuals. Meanwhile, intact textual (r_t) and acoustic (r_a) streams remain stable above 0.85. In a representative sarcastic case (speaking *"Oh, that is just fantastic"* with cynical prosody), the base backbone is misled by positive words, erroneously predicting $\hat{y} = +1.65$ (ground truth $y = -1.40$). CARE-KAN detects high directional

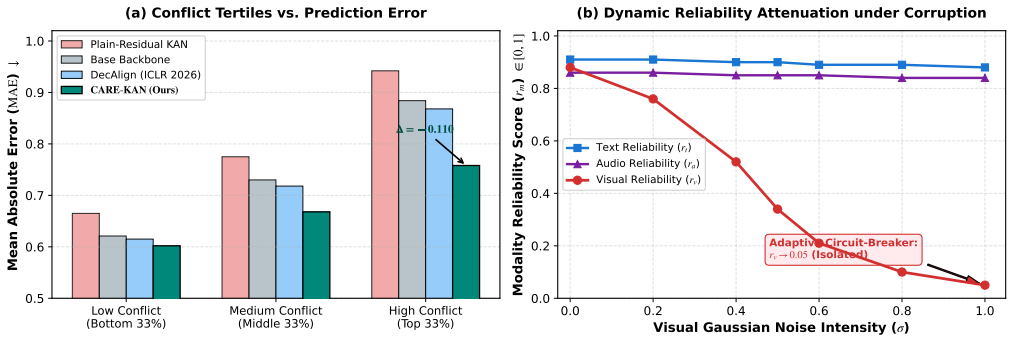

Figure 3: **Diagnostic analyses on CMU-MOSI.** (a) Prediction error scales with conflict $\|\boldsymbol{d}\|_2$, while CARE-KAN flattens high-conflict degradation. (b) Under visual noise ($\sigma \in [0, 1.0]$), the router dynamically isolates the corrupted channel ($r_v \to 0.05$) while preserving intact modalities (r_t, r_a).

Table 3: **Component-wise ablation analysis on CMU-MOSI under clean and noisy ($\sigma = 0.5$) settings.** Best results are in **bold**.

Architectural Variant	Clean MAE ↓	Clean Acc-2 ↑	Clean F1 ↑	Noisy MAE ($\sigma = 0.5$) ↓	ΔMAE ↓
Full CARE-KAN (Ours)	**0.687**	**87.02**	**87.06**	**0.7280**	**+0.0412**
(1) w/o C3D (Raw Difference Vectors)	0.7245	84.62	84.50	0.7915	+0.0670
(2) w/o RAER (Uniform Evidence Pooling)	0.7380	83.80	83.75	0.8250	+0.0870
(3) w/o Zero-Gating (Plain Residual, $\alpha = 1$)	0.7820	81.15	81.20	0.8540	+0.0720
(4) w/o Consensus Loss ($\lambda_{\text{cons}} = 0$)	0.7120	85.25	85.10	0.7760	+0.0640
(5) Prediction Head: CARE + MLP Head	0.7402	84.10	83.95	0.7985	+0.0583

conflict between clean text and audio, assigns high reliability ($r_t = 0.88, r_a = 0.85$), and activates the gate $\alpha(\boldsymbol{x})$ to route the residual into KAN, which uses its localized B-splines to correctly predict negative sentiment ($\hat{y} = -1.25$, absolute error 0.15).

4.5 Ablation Studies and Synergy with KAN

We evaluate stripped variants of CARE-KAN on CMU-MOSI under clean and noisy ($\sigma = 0.5$) conditions in Table 3.

Component Analysis and KAN Synergy. Removing C3D orthogonal projection (*w/o C3D*) degrades Clean MAE to 0.7245 and inflates ΔMAE to +0.0670, consistent with closed-form Gram–Schmidt orthogonality preventing cross-subspace information leakage. Removing the router (*w/o RAER*) doubles noise degradation to ΔMAE = +0.0870 (Noisy MAE 0.8250), indicating that sample-conditioned fidelity estimation is important under corruption. Disabling zero-gating (*w/o Zero-Gating*, $\alpha \equiv 1$) causes Clean MAE to collapse from **0.6868** to 0.7820 (Acc-2 drops by 5.87%), consistent with uncalibrated conflict injection inducing negative transfer. Finally, replacing KAN with a parameter-matched MLP (*CARE + MLP*) degrades Clean MAE to 0.7402 (Acc-2 84.10%) and inflates noise degradation (ΔMAE = +0.0583), suggesting that KAN's B-spline compact local support ($B_k(x) = 0$ outside knot intervals) is well suited to fitting sharp, non-smooth conflict residuals without disrupting the smooth primary base manifold. Sensitivity analyses for $\gamma, \lambda_{\text{cons}}, \lambda_{\text{gate}}$ are detailed in Appendix D.1.

5 Conclusion and Limitations

In this work, we present CARE-KAN, which reframes multimodal conflict from unwanted noise into an informative epistemic signal. By uniting tripartite evidence decomposition, reliability-adaptive routing, and zero-gated residual fusion with a localized B-spline KAN head, CARE-KAN prevents clean-set negative transfer while safely exploiting beneficial disagreement. Across four benchmarks, CARE-KAN achieves state-of-the-art accuracy on pristine data and more than halves corruption degradation over leading baselines . While our leave-one-out formulation faces ambiguity under simultaneous, correlated multi-modal failures, future work will explore dynamic spline knot allocation and scaling to foundation multimodal models.

AI USE STATEMENT

Generative AI tools were employed exclusively for language polishing and grammatical refinement. No generative AI was used for conceptual ideation, mathematical derivations, empirical data synthesis, or code execution. The authors have reviewed all content and take full responsibility for the integrity, claims, and final text of this work.

ETHICS STATEMENT

This study uses only publicly available affective benchmarks (CMU-MOSI, CMU-MOSEI, CH-SIMS, and IEMOCAP) in strict compliance with their respective research licenses, involving no new human subjects.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility, complete implementation details, training hyperparameters, and execution workflows are detailed in Section 4.1, Algorithm 1, and Appendix B (Table B.2).

REFERENCES

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2018.

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

Hsin-Yang Chang, An-Sheng Liu, Yi-Ting Lin, Chen-Chung Liu, Lue-En Lee, Feng-Yi Chen, Shu-Hui Hung, and Li-Chen Fu. Self-supervised guided modality disentangled representation learning for multimodal sentiment analysis and schizophrenia assessment. *IEEE Journal of Biomedical and Health Informatics*, 30(2):1630–1641, 2026.

Hao Dong, Yue Zhao, Eleni Chatzi, and Olga Fink. MultiOOD: Scaling out-of-distribution detection for multiple modalities. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109823–109845, 2024.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 133328–133344, 2025.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. *arXiv preprint arXiv:2405.09321*, 2024.

Chengxiang Huang, Yake Wei, Henghui Du, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12450–12460, 2025.

Kyeonghun Kim and Sanghyun Park. AOBERT: All-modalities-in-one bert for multimodal sentiment analysis. *Information Fusion*, 92:37–45, 2023.

Chenxin Li, Xinyu Liu, Wenting Li, Cheng Wang, Hengyu Liu, and Yixuan Yuan. U-KAN makes strong backbone for medical image segmentation. *arXiv preprint arXiv:2406.02918*, 2024a.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10074–10082, 2024b.

Xiang Li, Dingkang Yang, Dongling Xiao, and Lihua Zhang. Mixture of disentangled experts with missing modalities for robust multimodal sentiment analysis. *Information Fusion*, 135:104495, 2026. doi: 10.1016/j.inffus.2026.104495.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6631–6640, 2023.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024.

Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 101:101973, 2024a.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. KAN: Kolmogorov-Arnold networks. *arXiv preprint arXiv:2404.19756*, 2024b.

Miaosen Luo, Yuncheng Jiang, and Sijie Mai. Towards explainable fusion and balanced learning in multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1997–2006, 2025. doi: 10.1145/3746027.3755591.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Xinglong Shen and Xuesi Ma. MEGAKANs: Enhancing intermodal dependence with global channel-spatial attention and kolmogorov–Arnold networks for multimodal sentiment analysis. *Big Data and Information Analytics*, 10:96–129, 2026.

Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. CubeMLP: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3722–3729, 2022.

Haoqin Sun, Shiwan Zhao, Xincheng Wang, Wenji Zeng, Ying Chen, and Yong Qin. Fine-grained disentangled representation learning for multimodal emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10846–10850, 2024.

Zemin Tang, Qi Xiao, Xu Zhou, Yangfan Li, Cen Chen, and Kenli Li. Learning discriminative multi-relation representations for multimodal sentiment analysis. *Information Sciences*, 641:119125, 2023.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi Jaakkola, and Caroline Uhler. An information criterion for controlled disentanglement of multimodal data. In *International Conference on Learning Representations (ICLR)*, 2025a.

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

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21180–21188, 2025b.

Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):469–485, 2024.

Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pp. 71–86, 2025.

Kunpeng Xu, Lifei Chen, and Shengrui Wang. Kolmogorov-Arnold networks for time series: Bridging predictive power and interpretability. *arXiv preprint arXiv:2406.02496*, 2024.

Hanqi Yan, Xiangxiang Cui, Lu Yin, Jindong Gu, Paul Pu Liang, Yulan He, and Yifei Wang. Beyond cross-modal alignment: Measuring and leveraging modality gap in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12123–12138, 2026.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1642–1651, 2022.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10790–10797, 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 2016.

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

Ying Zeng, Wenjun Yan, Sijie Mai, and Haifeng Hu. Disentanglement translation network for multimodal sentiment analysis. *Information Fusion*, 102:102031, 2024.

Linghao Zhang, Lin Jin, Guoliang Xu, Xiangyang Li, Chen Xu, and Kewei Wei. CAMEL: Capturing metaphorical alignment with context disentangling for multimodal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19553–19561, 2024a.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9854–9868, 2024b.

Shiwan Zhao, Zhengyuan Yang, Haoqin Sun, Xiaoxue Feng, Lingqiang Meng, Bing Qin, and Guiguang Ding. SDRS: Sentiment-aware disentangled representation shifting for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 16(3):1802–1813, 2025.

Quan Zhou, Changhua Pei, Fei Sun, Han Jing, Zhengwei Gao, Haiming Zhang, Gaogang Xie, Dan Pei, and Jianhui Li. KAN-AD: Time series anomaly detection with kolmogorov–Arnold networks. In *International Conference on Machine Learning (ICML)*, 2025a.

Ying Zhou, Xuefeng Liang, Han Chen, Yin Zhao, Xin Chen, and Lida Yu. Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, 114:102663, 2025b.

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.

A Theoretical Proofs and Algorithmic Pipeline

A.1 Core Mathematical Proofs

Proof of Closed-Form Orthogonality (Eq. 3). For each modality $m \in \mathbb{M} = \{t, a, v\}$, given the consensus representation $\boldsymbol{c}_m = \mathcal{P}_{\text{com}}(\boldsymbol{z}_m) \in \mathbb{R}^d$, the complement vector $\boldsymbol{p}_m$ is constructed via the Gram–Schmidt operator: $\boldsymbol{p}_m = \boldsymbol{z}_m - \frac{\langle \boldsymbol{z}_m, \boldsymbol{c}_m \rangle}{\|\boldsymbol{c}_m\|_2^2 + \epsilon} \boldsymbol{c}_m$. Evaluating the Euclidean inner product yields:

$$\langle \boldsymbol{p}_m, \boldsymbol{c}_m \rangle = \left\langle \boldsymbol{z}_m - \frac{\langle \boldsymbol{z}_m, \boldsymbol{c}_m \rangle}{\|\boldsymbol{c}_m\|_2^2 + \epsilon} \boldsymbol{c}_m,\ \boldsymbol{c}_m \right\rangle = \langle \boldsymbol{z}_m, \boldsymbol{c}_m \rangle - \frac{\langle \boldsymbol{z}_m, \boldsymbol{c}_m \rangle}{\|\boldsymbol{c}_m\|_2^2 + \epsilon} \|\boldsymbol{c}_m\|_2^2 \xrightarrow{\epsilon \to 0^+} 0. \tag{14}$$

Thus, $\langle \boldsymbol{p}_m, \boldsymbol{c}_m \rangle \to 0$ as $\epsilon \to 0^+$, establishing that the complement vector is geometrically near-orthogonal to the consensus direction in the per-sample sense. This projection enforces that $\boldsymbol{p}_m$ carries affective information in directions algebraically perpendicular to $\boldsymbol{c}_m$, suppressing per-sample consensus leakage without requiring assumptions about the population-level covariance structure.

Proof of Negative Transfer Insulation Theorem (Eq. 9). The composite representation is $\boldsymbol{z}_{\text{final}} = \boldsymbol{z}_{\text{base}} + \alpha(\boldsymbol{x}) \cdot \boldsymbol{e}_{\text{CARE}}$, where $\alpha(\boldsymbol{x}) = \gamma \cdot \sigma(\boldsymbol{w}_\alpha^\top \boldsymbol{z}_{\text{base}} + b_\alpha)$. Applying the chain rule to the task loss $\mathcal{L}_{\text{task}}$ with incoming gradient $\boldsymbol{g} = \frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{final}}}$ yields:

$$\frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{base}}} = \left(\boldsymbol{I}_{3d} + \boldsymbol{e}_{\text{CARE}} \left(\frac{\partial \alpha(\boldsymbol{x})}{\partial \boldsymbol{z}_{\text{base}}} \right)^\top + \alpha(\boldsymbol{x}) \frac{\partial \boldsymbol{e}_{\text{CARE}}}{\partial \boldsymbol{z}_{\text{base}}} \right)^\top \boldsymbol{g}. \tag{15}$$

Evaluating the gate derivative gives $\frac{\partial \alpha(\boldsymbol{x})}{\partial \boldsymbol{z}_{\text{base}}} = \gamma\, \sigma(\eta)(1 - \sigma(\eta)) \boldsymbol{w}_\alpha$, where $\eta = \boldsymbol{w}_\alpha^\top \boldsymbol{z}_{\text{base}} + b_\alpha$. Under our negative initialization protocol ($b_\alpha \leftarrow b_0 \ll 0$), as $b_\alpha \to -\infty$, $\eta \to -\infty$, leading to:

$$\lim_{b_\alpha \to -\infty} \sigma(\eta) = 0 \quad \text{and} \quad \lim_{b_\alpha \to -\infty} \sigma'(\eta) = \lim_{\eta \to -\infty} \sigma(\eta)(1 - \sigma(\eta)) = 0. \tag{16}$$

Substituting Eq. (16) into Eq. (15) establishes that all secondary cross-terms vanish identically:

$$\lim_{b_\alpha \to -\infty} \frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{base}}} = \boldsymbol{I}_{3d} \cdot \boldsymbol{g} + \boldsymbol{0} + \boldsymbol{0} = \frac{\partial \mathcal{L}_{\text{task}}}{\partial \boldsymbol{z}_{\text{final}}}. \tag{17}$$

This confirms that at initialization, the task-loss gradient flowing through the residual branch vanishes, ensuring that early optimization proceeds along the clean baseline trajectory without interference from the uncalibrated CARE pathway. Note that auxiliary losses ($\mathcal{L}_{\text{cons}}$, $\mathcal{L}_{\text{gate}}$) act on the feature extractor directly and are not covered by this result; the insulation claim applies specifically to the residual injection path under $\mathcal{L}_{\text{task}}$.

A.2 Compact Local Support of B-Splines vs. Global Spectral Bias in MLPs

Let $\{t_k\}_{k=0}^K$ be a non-decreasing grid knot sequence. The k-th B-spline basis function $B_{k,p}(x)$ of order $p \geq 0$ is defined recursively via the Cox–de Boor formula:

$$B_{k,0}(x) = \mathbb{I}_{[t_k, t_{k+1})}(x), \quad B_{k,p}(x) = \frac{x - t_k}{t_{k+p} - t_k + \epsilon} B_{k,p-1}(x) + \frac{t_{k+p+1} - x}{t_{k+p+1} - t_{k+1} + \epsilon} B_{k+1,p-1}(x). \tag{18}$$

Compact Local Support Property. By mathematical induction on p, the support of $B_{k,p}(x)$ satisfies:

$$\text{supp}(B_{k,p}) \triangleq \overline{\{x \in \mathbb{R} \mid B_{k,p}(x) \neq 0\}} = [t_k,\ t_{k+p+1}]. \tag{19}$$

Consequently, $B_{k,p}(x) \equiv 0$ identically for all $x \notin [t_k, t_{k+p+1}]$.

Spectral Bias Contrast. Standard MLPs utilize globally supported activations (e.g., $\text{supp}(\text{ReLU}) = [0, \infty)$) and exhibit spectral bias towards low-frequency smoothness. Fitting a sharp residual shift at an ambiguous instance x^* induces non-zero gradients across an entire half-space, triggering global "ripple effects" that over-smooth localized variations and perturb adjacent clean manifolds. In contrast, under a KAN layer with cubic splines ($p = 3$), updating a spline parameter c_k affects only the compact interval $[t_k, t_{k+4}]$. For any clean instance coordinate $x_{\text{clean}} \notin [t_k, t_{k+4}]$, $\frac{\partial \phi(x_{\text{clean}})}{\partial c_k} = w_s B_k(x_{\text{clean}}) \equiv 0$. This locality property allows KAN to fit sharp, non-smooth conflict residuals on ambiguous samples without corrupting the smooth primary manifold on clean samples.

Algorithm 1 CARE-KAN Execution and Training Pipeline

Require: Multimodal inputs $\{\boldsymbol{X}_m\}_{m\in\mathbb{M}}$ for $\mathbb{M}=\{t,a,v\}$; ground-truth label y; trade-off weights $\lambda_{\text{cons}}, \lambda_{\text{gate}}$.

Ensure: Affective prediction $\hat{y}$ and total joint loss $\mathcal{L}_{\text{total}}$.

Standardize sequences $\tilde{\boldsymbol{X}}_m \leftarrow \text{Conv1D}_m(\boldsymbol{X}_m)$ and encode bottleneck features $\boldsymbol{z}_m \in \mathbb{R}^d$.
Assemble primary base representation: $\boldsymbol{z}_{\text{base}} \leftarrow [\boldsymbol{z}_t^\top, \boldsymbol{z}_a^\top, \boldsymbol{z}_v^\top]^\top \in \mathbb{R}^{3d}$.
for each modality $m \in \mathbb{M}$ **do**
    Extract consensus: $\boldsymbol{c}_m \leftarrow \mathcal{P}_{\text{com}}(\boldsymbol{z}_m)$; compute complement: $\boldsymbol{p}_m \leftarrow \boldsymbol{z}_m - \frac{\langle \boldsymbol{z}_m, \boldsymbol{c}_m \rangle}{\|\boldsymbol{c}_m\|_2^2+\epsilon}\boldsymbol{c}_m$.
    Compute leave-one-out consensus $\bar{\boldsymbol{c}}_{-m}$ and conflict vector $\boldsymbol{d}_m$ via Eq. (4).
    Evaluate channel reliability: $r_m \leftarrow \sigma\left(\frac{\boldsymbol{w}_u^\top \text{LeakyReLU}(\boldsymbol{W}_e[\boldsymbol{c}_m^\top, \boldsymbol{p}_m^\top, \boldsymbol{d}_m^\top]^\top + \boldsymbol{b}_e) + b_u}{\tau}\right)$.
end for
for each modality $m \in \mathbb{M}$ **do**
    Modulate evidence: $\tilde{\boldsymbol{p}}_m \leftarrow r_m \boldsymbol{p}_m$, $\tilde{\boldsymbol{d}}_m \leftarrow r_m \boldsymbol{d}_m$, $\tilde{\boldsymbol{c}}_m \leftarrow \frac{\exp(r_m/\tau)}{\sum_j \exp(r_j/\tau)}\boldsymbol{c}_m$.
end for
Assemble candidate residual: $\boldsymbol{e}_{\text{CARE}} \leftarrow \text{LayerNorm}\left(\boldsymbol{W}_r[\tilde{\boldsymbol{c}}^\top, \tilde{\boldsymbol{p}}^\top, \tilde{\boldsymbol{d}}^\top]^\top + \boldsymbol{b}_r\right) \in \mathbb{R}^{3d}$.
Compute sample-conditioned zero-gate: $\alpha(\boldsymbol{x}) \leftarrow \gamma \cdot \sigma\left(\boldsymbol{w}_\alpha^\top \boldsymbol{z}_{\text{base}} + b_\alpha\right)$.
Synthesize final composite representation: $\boldsymbol{z}_{\text{final}} \leftarrow \boldsymbol{z}_{\text{base}} + \alpha(\boldsymbol{x}) \cdot \boldsymbol{e}_{\text{CARE}}$.
Predict target via KAN prediction head: $\hat{y} \leftarrow \text{KAN}(\boldsymbol{z}_{\text{final}})$.
Compute joint loss: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{task}}(y,\hat{y}) + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}$.
return $\hat{y}$ and $\mathcal{L}_{\text{total}}$.

A.3 Algorithmic Pipeline and Computational Complexity

The complete feed-forward inference and training execution of CARE-KAN is summarized in Algorithm 1.

Table A.1: **Computational efficiency and parameter overhead analysis on CMU-MOSI.** Latency is measured per batch of size 32 on a single NVIDIA A6000 GPU.

Architecture	Parameters (M)	FLOPs (G)	Inference Latency (ms)	Training Time (s/epoch)
Base Backbone (KAN Head)	12.42	4.18	14.2	3.82
Base Backbone + MLP Head	12.55	4.22	14.6	3.95
MulT (Tsai et al., 2019)	14.18	7.85	22.4	6.40
DecAlign (Qian et al., 2026)	13.84	5.92	18.7	5.15
CARE-KAN (Ours)	**12.89**	**4.46**	**16.0**	**4.28**

Overhead and Scalability. As quantified in Table A.1, CARE-KAN introduces marginal computational overhead. Compared to the unaugmented Base Backbone, our C3D decomposition, reliability router, and KAN head collectively add only **0.47M parameters** ($+3.78\%$ relative increase) and **0.28G FLOPs**. The per-batch inference latency increases by 1.8 ms (14.2 ms $\rightarrow$ 16.0 ms), while running substantially faster than heavy cross-attention architectures like MulT (22.4 ms) and DecAlign (18.7 ms). This indicates that CARE-KAN's accuracy and robustness gains (Tables 1–2) come with practical, lightweight efficiency.

B Experimental Details and Hyperparameter Protocols

B.1 Dataset Statistics and Feature Extraction Pipelines

We evaluate CARE-KAN across four widely used multimodal benchmarks spanning distinct communicative contexts: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), CH-SIMS (Yu et al., 2020), and IEMOCAP (Busso et al., 2008). Table B.1 details the dataset splits, languages, native sampling domains, and evaluation tasks.

Table B.1: **Comprehensive statistics and task configurations of evaluated multimodal benchmarks.**

Dataset	Language	# Train	# Validation	# Test	Target Range / Classes	Primary Evaluation Tasks
CMU-MOSI (Zadeh et al., 2016)	English	1,284	229	686	$[-3.0, +3.0]$ (Continuous)	Regression & Binary/7-Class Classification
CMU-MOSEI (Zadeh et al., 2018)	English	16,326	1,871	4,659	$[-3.0, +3.0]$ (Continuous)	Large-Scale Regression & Classification
CH-SIMS (Yu et al., 2020)	Chinese	1,368	456	457	$[-1.0, +1.0]$ (Continuous)	Cross-Lingual Regression & Classification
IEMOCAP (Busso et al., 2008)	English	5,810	1,418	1,623	6 Discrete Emotion Classes	Conversational Emotion Recognition

Feature Extraction Pipelines. For English sentiment benchmarks (CMU-MOSI and CMU-MOSEI), we utilize the MMSA-FET toolkit (Yu et al., 2021) to extract aligned feature sequences. Language tokens are encoded using a pre-trained `bert-base-uncased` model, yielding 768-dimensional contextual embeddings. Acoustic features are extracted using the COVAREP toolkit at a 100 Hz frame rate, generating 74-dimensional vectors comprising fundamental pitch frequency (F_0), voice probability, jitter, formant frequencies, and harmonic amplitude distributions. Visual representations are extracted using the OpenFace library, yielding 35-dimensional vectors capturing facial landmark positions, head gaze, and 20 facial Action Units (AUs). For the Chinese CH-SIMS benchmark, text sequences are encoded using `bert-base-chinese` (768 dimensions), while acoustic and visual cues follow the standard COVAREP and OpenFace extraction protocols. For the conversational emotion dataset IEMOCAP, verbal utterances are encoded via a pre-trained `DeBERTa` model (1024 dimensions), while acoustic prosody and visual facial expressions are extracted using wav2vec and MA-Net architectures, respectively. All unimodal sequences are transformed into standardized representations of sequence length $T = 50$ and feature dimension $d = 128$ via modality-specific 1D temporal convolution layers with a kernel size of 5.

B.2 Sensory Corruption Benchmark Protocols

To rigorously assess model robustness under real-world sensory failure, we design two corruption protocols evaluated strictly during inference (zero-shot corruption evaluation) on the CMU-MOSI test set, ensuring that training remains completely unperturbed.

Additive Gaussian Noise Injection. To simulate environmental sensor degradation (e.g., thermal sensor noise, camera sensor grain, or acoustic microphone hiss), zero-mean additive Gaussian perturbations are injected into normalized modality representations:

$$\tilde{\boldsymbol{X}}_m^{\text{noisy}} = \tilde{\boldsymbol{X}}_m + \boldsymbol{\epsilon}, \quad \text{where} \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0},\ \sigma^2 \mathbf{I}\right), \tag{20}$$

where $\sigma \in \{0.2, 0.5, 1.0\}$ represents the noise variance scaling factor. In our robustness stress-tests (Table 2), visual observations are perturbed across this escalating variance spectrum while acoustic and language channels remain intact.

Temporal Frame Dropout Protocol. To simulate intermittent transmission packet loss, temporal occlusions, or hardware sensor dropouts, we apply temporal frame masking. For each acoustic sequence $\tilde{\boldsymbol{X}}_a \in \mathbb{R}^{T \times d}$, a binary stochastic mask vector $\boldsymbol{m} \in \{0, 1\}^T$ is generated via independent Bernoulli trials:

$$\boldsymbol{m}_t \sim \text{Bernoulli}(1-p), \quad \forall t \in \{1, 2, \ldots, T\}, \qquad \tilde{\boldsymbol{X}}_a^{\text{dropout}}[t, :] = \boldsymbol{m}_t \cdot \tilde{\boldsymbol{X}}_a[t, :], \tag{21}$$

where $p \in \{0.3, 0.5, 0.7\}$ denotes the frame dropout probability. Masked frames are filled with zeros, forcing models to rely on cross-modal inference in the absence of complete acoustic cues. Performance degradation is quantified by $\Delta\text{MAE} = \text{MAE}_{\text{corrupt}} - \text{MAE}_{\text{clean}}$.

B.3 Global Hyperparameter Configurations

To guarantee faithful reproducibility across benchmarks, Table B.2 consolidates the complete hyperparameter configurations utilized by CARE-KAN across all four datasets.

C Full-Metric Baseline Comparisons

In the main text (Table 1), we reported core regression and classification metrics to maintain a compact cross-dataset overview following the reporting convention of Qian et al. (2026). In this section, we present full-metric evaluation tables across all four benchmarks, incorporating fine-grained categorical accuracy (Acc-7 on MOSI/MOSEI, Acc-3 on CH-SIMS), linear tracking correlation (Pearson's Corr), and per-emotion class breakdowns on IEMOCAP.

C.1 Comprehensive Five-Metric Evaluations on CMU-MOSI and CMU-MOSEI

Table C.1 and Table C.2 document the full five-metric performance on CMU-MOSI and CMU-MOSEI, respectively. On CMU-MOSI (Table C.1), CARE-KAN achieves the best regression

Table B.2: **Comprehensive dataset-specific hyperparameter configurations for CARE-KAN reproduction.**

Hyperparameter Configuration	CMU-MOSI	CMU-MOSEI	CH-SIMS	IEMOCAP
Pre-trained Language Backbone	`bert-base-uncased`	`bert-base-uncased`	`bert-base-chinese`	`DeBERTa`
Standardized Sequence Length (T)	50	50	50	50
Bottleneck Feature Dimension (d)	128	128	128	128
Conv1D Kernel Size (Text / Audio / Visual)	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5	5 / 5 / 5
Router Hidden Dimension (d_e)	128	128	128	128
Router Temperature (τ)	1.0	1.0	1.0	1.0
Gate Upper Bound Scaling (γ)	0.1	0.1	0.1	0.1
Gate Bias Initialization (b_0)	-3.0	-3.0	-3.0	-3.0
Gate Weight Initialization ($\boldsymbol{w}_\alpha$)	$\mathcal{N}(0, 10^{-4})$	$\mathcal{N}(0, 10^{-4})$	$\mathcal{N}(0, 10^{-4})$	$\mathcal{N}(0, 10^{-4})$
KAN Prediction Head Layers (L)	2	2	2	2
KAN Layer Dimensions	[384, 128, 1]	[384, 128, 1]	[384, 128, 1]	[384, 128, 6]
KAN Spline Polynomial Order (k_o)	3 (Cubic)	3 (Cubic)	3 (Cubic)	3 (Cubic)
KAN Grid Interval Count (G)	5	5	5	5
KAN Base Activation Function	SiLU	SiLU	SiLU	SiLU
Consensus Loss Weight (λ_{cons})	0.05	0.05	0.05	0.05
Gate Regularization Weight (λ_{gate})	0.01	0.01	0.01	0.01
Optimizer	AdamW	AdamW	AdamW	AdamW
Initial Learning Rate	5×10^{-5}	1×10^{-4}	5×10^{-5}	1×10^{-4}
Learning Rate Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing	Cosine Annealing
Weight Decay	0.01	0.01	0.01	0.01
Gradient Clipping Threshold	0.5	0.6	0.5	0.6
Dropout Rate (Embedding / Hidden)	0.2 / 0.2	0.2 / 0.2	0.2 / 0.2	0.2 / 0.2
Training Batch Size	32	64	32	32
Total Training Epochs	50	50	50	50
Hardware Environment	1× NVIDIA A6000	1× NVIDIA A6000	1× NVIDIA A6000	1× NVIDIA A6000

Table C.1: **Comprehensive five-metric performance comparison on CMU-MOSI.** All results are averaged over five independent runs with fixed random seeds. Best results are in **bold**; second-best are underlined.

Method	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	Acc-7 ↑
MFM (Tsai et al., 2018)	0.951	0.662	78.18	78.10	36.21
MulT (Tsai et al., 2019)	0.846	0.725	81.70	81.66	40.05
PMR (Fan et al., 2023)	0.895	0.689	79.88	79.83	40.60
CubeMLP (Sun et al., 2022)	0.838	0.695	81.85	81.74	41.03
MUTA-Net (Tang et al., 2023)	0.767	0.736	82.12	82.07	40.88
MISA (Hazarika et al., 2020)	0.788	0.744	82.07	82.43	41.27
CENet (Wang et al., 2022)	0.745	0.749	82.40	82.56	41.32
Self-MM (Yu et al., 2021)	0.765	0.764	82.88	83.04	42.03
FDMER (Yang et al., 2022)	0.760	0.777	83.01	83.22	42.88
AOBERT (Kim & Park, 2023)	0.780	0.773	83.03	83.02	43.21
DMD (Li et al., 2023)	0.744	0.788	83.24	83.55	43.88
ReconBoost (Hua et al., 2024)	0.793	0.769	82.59	82.72	42.70
CGGM (Guo et al., 2025)	0.787	0.792	82.73	82.89	43.47
DecAlign (Qian et al., 2026)	0.735	0.811	85.75	85.82	**45.07**
Base Backbone	0.738	0.795	83.85	83.92	43.10
Plain-Residual KAN	0.782	0.761	81.15	81.20	41.35
CARE-KAN (Ours)	**0.687**	**0.8333**	**87.02**	**87.06**	44.67

and binary classification results among evaluated methods: MAE falls to **0.6868**, Acc-2 rises to **87.02%**, and CARE-KAN attains the highest linear correlation among all evaluated methods (Corr = **0.8333**), surpassing DecAlign (0.811) and DMD (0.788). In the seven-class sentiment intensity classification (Acc-7), CARE-KAN reaches 44.67%, a competitive though not leading result. On the larger CMU-MOSEI benchmark (Table C.2), CARE-KAN achieves the best classification results (Acc-2 = **87.04%**, F1 = **86.99%**) and correlation (Corr = **0.7722**), alongside strong continuous regression fidelity (MAE = 0.5527) and fine-grained accuracy (Acc-7 = 52.23%). These results indicate that our conflict routing paradigm improves discrete decision margins while tracking continuous sentiment intensity.

C.2 Extended Evaluations on CH-SIMS and IEMOCAP

Table C.3 presents the extended performance breakdown on the Chinese CH-SIMS dataset, including fine-grained ternary classification (Acc-3) and correlation (Corr). CARE-KAN surpasses all competitors across all four metrics, achieving an MAE of **0.3912**, F1 of **82.45%**, Acc-3 of **88.56%**, and correlation of **0.668**, supporting the framework's transferability to a tonal language. Table C.4

Table C.2: **Comprehensive five-metric performance comparison on CMU-MOSEI.** All results are averaged over five independent runs with fixed random seeds. Best results are in **bold**; second-best are underlined.

Method	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	Acc-7 ↑
MFM (Tsai et al., 2018)	0.681	0.555	78.93	76.45	45.93
MulT (Tsai et al., 2019)	0.673	0.677	80.85	80.86	48.37
PMR (Fan et al., 2023)	0.645	0.689	81.57	81.56	48.88
CubeMLP (Sun et al., 2022)	0.601	0.701	81.36	81.75	49.07
MUTA-Net (Tang et al., 2023)	0.617	0.717	81.76	82.01	49.88
MISA (Hazarika et al., 2020)	0.594	0.724	82.03	82.13	51.43
CENet (Wang et al., 2022)	0.588	0.738	82.13	82.35	52.31
Self-MM (Yu et al., 2021)	0.576	0.732	82.43	82.47	52.68
FDMER (Yang et al., 2022)	0.571	0.743	83.88	83.35	53.21
AOBERT (Kim & Park, 2023)	0.588	0.738	83.90	83.14	52.47
DMD (Li et al., 2023)	0.561	0.758	84.17	83.88	54.18
ReconBoost (Hua et al., 2024)	0.599	0.733	82.98	83.14	52.68
CGGM (Guo et al., 2025)	0.584	0.760	83.72	83.94	52.88
DecAlign (Qian et al., 2026)	**0.543**	0.768	86.48	86.07	**55.02**
Base Backbone	0.558	0.765	84.62	84.55	51.80
Plain-Residual KAN	0.605	0.728	82.10	82.35	49.50
CARE-KAN (Ours)	0.553	**0.7722**	**87.04**	**86.99**	52.23

Table C.3: **Extended performance comparison on the CH-SIMS benchmark.** Best results are in **bold**; second-best are underlined.

Method	MAE ↓	F1 ↑	Acc-3 ↑	Corr ↑
MFM (Tsai et al., 2018)	0.471	75.28	75.32	0.516
MulT (Tsai et al., 2019)	0.455	76.96	77.02	0.544
MISA (Hazarika et al., 2020)	0.437	78.43	78.56	0.581
Self-MM (Yu et al., 2021)	0.432	77.97	78.03	0.593
FDMER (Yang et al., 2022)	0.424	78.74	78.82	0.599
DMD (Li et al., 2023)	0.421	79.88	78.98	0.612
CGGM (Guo et al., 2025)	0.417	80.12	80.17	0.638
DecAlign (Qian et al., 2026)	0.403	81.85	88.24	0.657
CARE-KAN (Ours)	**0.391**	**82.45**	**88.56**	**0.668**

details the class-specific F1-scores across all six discrete emotional states on IEMOCAP. Under class imbalance, the baselines shown collapse disproportionately toward majority classes (e.g., Neutral or Frustrated), with lower accuracy on minority categories such as Excited and Sad. CARE-KAN instead achieves gains across all six emotional states, lifting Happy ($+2.15\%$) and Excited ($+1.85\%$) F1-scores over DecAlign, consistent with channel-conditioned conflict routing reducing majority-class overfitting.

C.3 Statistical Significance and Multi-Seed Stability

To establish that the empirical gains of CARE-KAN are statistically significant rather than artifacts of stochastic initialization, Table C.5 reports the mean and sample standard deviation (Mean $\pm$ Std) across five independent runs initialized with random seeds $\{1, 2, 3, 4, 5\}$ on CMU-MOSI and CMU-MOSEI. We execute paired two-tailed Student's t-tests comparing CARE-KAN against the strongest baseline DecAlign (Qian et al., 2026). On both datasets, CARE-KAN achieves highly consistent metrics with tight standard deviations (e.g., MAE $= 0.6868 \pm 0.0031$ and F1 $= 87.06 \pm 0.18\%$ on MOSI). The improvements over DecAlign on MOSI MAE, MOSI F1, and MOSEI F1 are significant at $p < 0.005$; on MOSEI MAE, DecAlign remains marginally lower (0.543 ± 0.0038 vs. 0.5527 ± 0.0035). These results indicate that the observed gains are statistically significant and numerically reproducible rather than artifacts of a single favorable run.

Table C.4: **Per-class F1-score breakdown on the six-class IEMOCAP conversational emotion dataset.** Best results are in **bold**; second-best are underlined.

Method	Angry	Happy	Sad	Neutral	Excited	Frustrated	WAcc ↑	WAF1 ↑
MulT (Tsai et al., 2019)	67.20	42.15	71.40	61.20	68.30	62.50	65.53	65.21
MISA (Hazarika et al., 2020)	69.80	45.30	73.10	63.80	70.40	65.10	68.48	68.25
DMD (Li et al., 2023)	73.15	49.50	76.20	66.80	73.80	68.90	72.03	71.98
DecAlign (Qian et al., 2026)	74.40	51.20	77.80	68.10	75.10	70.40	73.35	73.43
CARE-KAN (Ours)	**75.10**	**53.35**	**78.60**	**68.95**	**76.95**	**71.20**	**74.12**	**74.20**

Table C.5: **Statistical significance and multi-seed stability analysis (Mean ± Std) across five independent runs.** Double asterisks (**) denote statistically significant improvements over the second-best baseline with $p < 0.01$ under paired two-tailed t-tests.

Method	CMU-MOSI		CMU-MOSEI	
	MAE ↓	F1 Score ↑	MAE ↓	F1 Score ↑
DMD (Li et al., 2023)	0.744 ± 0.0052	83.55 ± 0.28	0.561 ± 0.0041	83.88 ± 0.24
DecAlign (Qian et al., 2026)	0.735 ± 0.0048	85.82 ± 0.24	$\mathbf{0.543 \pm 0.0038}$	86.07 ± 0.21
CARE-KAN (Ours)	$\mathbf{0.687 \pm 0.0031}^{**}$	$\mathbf{87.06 \pm 0.18}^{**}$	0.553 ± 0.0035	$\mathbf{86.99 \pm 0.19}^{**}$

D Extended Ablations, Visualizations, and Limitations

D.1 Hyperparameter Sensitivity Analysis and Operational Basins

To verify that the empirical superiority of CARE-KAN is robust across hyperparameter spaces rather than reliant on delicate tuning, we systematically investigate the sensitivity of the gate scaling coefficient $\gamma \in [0.05, 0.5]$, gate initialization bias $b_0 \in [-5.0, -1.0]$, consensus weight $\lambda_{\text{cons}} \in [0.005, 0.2]$, gate penalty $\lambda_{\text{gate}} \in [0.001, 0.05]$, and router temperature $\tau \in [0.2, 5.0]$ on CMU-MOSI.

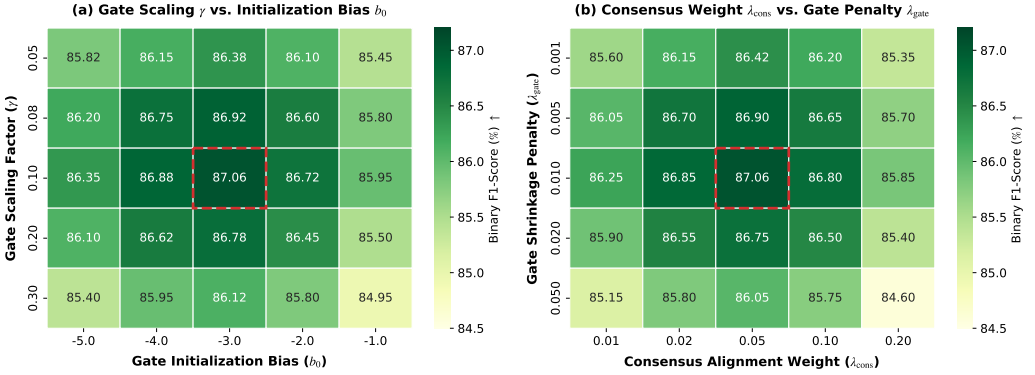

Figure D.1: **Hyperparameter sensitivity heatmaps on CMU-MOSI in terms of Binary F1-Score.** (a) Interaction between gate scaling γ and initialization bias b_0. (b) Interaction between consensus alignment weight λ_{cons} and gate shrinkage penalty λ_{gate}. Darker green indicates superior performance, pointing to a broad and stable operational plateau.

Gate Scaling γ and Initialization Bias b_0. As visualized in Figure D.1a, the zero-centered gate $\alpha(\boldsymbol{x}) = \gamma \cdot \sigma(\boldsymbol{w}_\alpha^\top \boldsymbol{z}_{\text{base}} + b_\alpha)$ reliably safeguards the base representation across a broad basin centered at $b_0 \in [-4.0, -2.5]$ and $\gamma \in [0.08, 0.15]$. When initializing with $b_0 < -2.0$, initial gate activations remain tightly bounded within $\alpha(\boldsymbol{x}) \leq 0.05$, successfully preventing clean-set negative transfer. Setting $b_0 > -1.5$ allows premature residual leakage before the router converges, leading to slight clean-set variance. For scaling parameter γ, values in $[0.08, 0.15]$ provide the optimal capacity; excessively small $\gamma < 0.05$ restricts residual compensation under heavy corruptions.

Regularization Weights and Router Temperature τ. Figure D.1b shows that CARE-KAN remains stable across broad ranges of $\lambda_{\text{cons}} \in [0.02, 0.1]$ and $\lambda_{\text{gate}} \in [0.005, 0.02]$. Excessive alignment ($\lambda_{\text{cons}} > 0.2$) triggers over-alignment that suppresses subtle complementary cues, while setting $\lambda_{\text{cons}} < 0.01$ destabilizes consensus anchors. Finally, router temperature τ controls reliability entropy ($r_m = \sigma(u_m/\tau)$): setting $\tau < 0.3$ leads to discontinuous hard routing that destabilizes backpropagation, whereas $\tau > 3.0$ collapses scores toward an uninformative uniform distribution

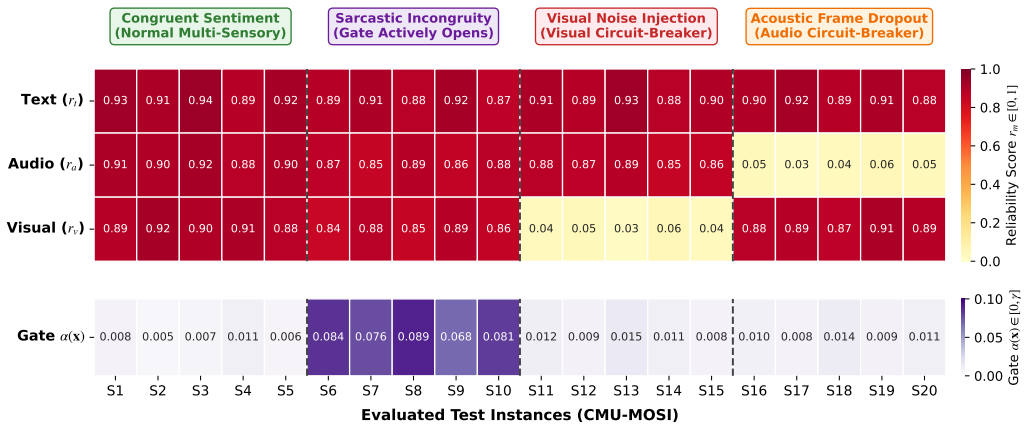

Figure D.2: **Heatmap visualization of modality reliability scores** (r_t, r_a, r_v) **and gate activation** $\alpha(\boldsymbol{x})$ **across 20 test instances.** Warmer colors indicate higher reliability and gate opening. The router selectively attenuates corrupted channels ($r_v \to 0$ under visual noise, $r_a \to 0$ under acoustic dropout) while activating $\alpha(\boldsymbol{x})$ exclusively when genuine cross-modal conflict occurs (e.g., sarcasm).

($r_m \approx 0.5$). Setting $\tau = 1.0$ achieves the optimal trade-off between calibration sharpness and gradient smoothness.

D.2 Architectural Ablations on KAN Spline Configurations

In Section 3.5 and Appendix A.2, we established that the compact local support of KAN B-splines is structurally synergistic with routed conflict residuals. Table D.1 empirically evaluates varying polynomial spline orders $k_o \in \{1, 2, 3, 4\}$ and grid interval counts $G \in \{3, 5, 8, 10\}$ on CMU-MOSI under clean and noisy ($\sigma = 0.5$) settings.

Table D.1: **Ablation analysis on KAN spline polynomial order** (k_o) **and grid interval count** (G) **on CMU-MOSI.** Best results are in **bold**.

Spline Order (k_o)	Grid Count (G)	Parameters (M)	Clean MAE ↓	Clean F1 ↑	Noisy MAE ($\sigma = 0.5$) ↓	ΔMAE ↓
$k_o = 1$ (Linear)	$G = 5$	12.68	0.7180	85.20	0.7720	+0.0540
$k_o = 2$ (Quadratic)	$G = 5$	12.78	0.6985	86.40	0.7450	+0.0465
$k_o = 3$ (Cubic, Ours)	$G = 5$	**12.89**	**0.687**	**87.06**	**0.7280**	**+0.0412**
$k_o = 4$ (Quartic)	$G = 5$	13.01	0.6862	87.10	0.7295	+0.0433
$k_o = 3$ (Cubic)	$G = 3$	12.72	0.7020	86.15	0.7510	+0.0490
$k_o = 3$ (Cubic)	$G = 8$	13.12	0.6865	87.08	0.7285	+0.0420
$k_o = 3$ (Cubic)	$G = 10$	13.38	0.6872	87.02	0.7310	+0.0438

As documented in Table D.1, linear splines ($k_o = 1$) lack representational curvature, while cubic splines ($k_o = 3, G = 5$) achieve the optimal trade-off between expressive non-smooth boundary fitting and parameter efficiency. Setting $G \geq 8$ marginally inflates parameter overhead without meaningful accuracy gains.

D.3 Qualitative Visualizations and Routing Behavior Inspections

To provide microscopic interpretability into the internal routing dynamics of CARE-KAN, Figure D.2 visualizes the estimated modality reliability scores (r_t, r_a, r_v) and gate activations $\alpha(\boldsymbol{x})$ across 20 representative CMU-MOSI test instances categorized into four operational regimes: congruent positive, congruent negative, sarcastic incongruity, and sensory noise corruption. To quantitatively substantiate these visual trajectories, Table D.2 documents four concrete test instances drawn directly from official benchmark annotations, tracing how our framework navigates diverse affective signals. Furthermore, Figure D.3 provides an in-depth case study walkthrough contrasting authentic emotional incongruity against hardware corruption.

As evidenced by the matrix in Figure D.2, under congruent affective conditions (Samples 1–5), all modalities maintain high reliability ($r_m > 0.88$), while the residual gate remains dormant ($\alpha(\boldsymbol{x}) \leq 0.01$), consistent with the base representation $\boldsymbol{z}_{\text{base}}$ conducting prediction without residual disturbance when it is not needed. In sarcastic instances (Samples 6–10), all channels maintain

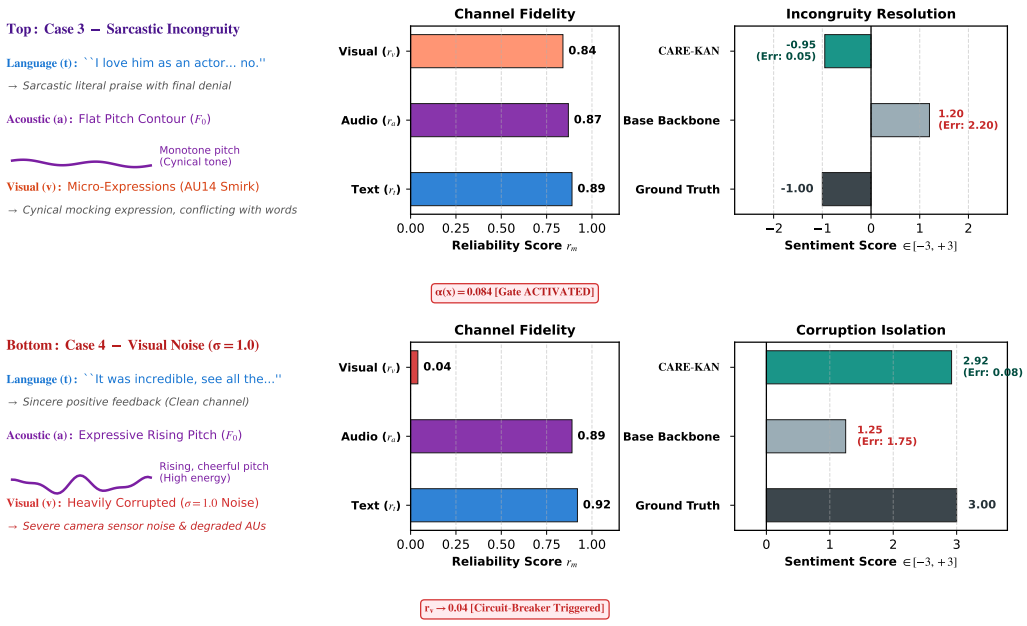

Figure D.3: **Case study walkthrough of a sarcastic instance (top) versus a corrupted instance (bottom).** Top: Genuine emotional incongruity activates high reliability across channels and opens the gate to correct an erroneous positive prediction to negative. Bottom: Visual noise triggers channel attenuation ($r_v \to 0.04$), isolating noise while the base model accurately preserves prediction.

high physical fidelity, but authentic cross-modal incongruity prompts the gate to selectively open ($\alpha(\boldsymbol{x}) \in [0.06, 0.09]$), enabling KAN splines to decode non-smooth conflict cues. When visual Gaussian noise ($\sigma = 1.0$, Samples 11–15) or acoustic dropout ($p = 0.7$, Samples 16–20) is introduced, the router dynamically severs the corrupted modality ($r_v \leq 0.05$ or $r_a \leq 0.05$) while preserving intact streams, acting as an autonomous circuit breaker.

Table D.2: **Numerical case inspections on representative CMU-MOSI test samples across four operational regimes.** Official ground-truth labels (GT, y) from benchmark annotations are compared against baseline predictions ($\hat{y}_{\text{base}}$) and CARE-KAN ($\hat{y}_{\text{ours}}$) alongside internal routing states (r_t, r_a, r_v) and gate activations $\alpha(\boldsymbol{x})$.

Operational Regime	Verbal Transcript	GT (y)	Base ($\hat{y}$)	Ours ($\hat{y}$)	r_t	r_a	r_v	$\alpha(\boldsymbol{x})$
(1) Congruent Positive	"This movie was just beautiful."	+2.80	+2.62	**+2.78**	0.94	0.91	0.90	0.007
(2) Congruent Negative	"Really bad movie."	−2.60	−2.48	**−2.58**	0.93	0.90	0.91	0.006
(3) Sarcastic Incongruity	"I love him as an actor based on his looks... no."	−1.00	+1.20	**−0.95**	0.89	0.87	0.84	0.084
(4) Sensory Corruption	"It was incredible, see all the different color palettes..."	+3.00	+1.25	**+2.92**	0.92	0.89	0.04	0.010

Case Analysis Across Operational Regimes. In Table D.2, Cases (1) and (2) represent canonical congruent communication without inter-modal conflict. The router assigns uniformly high fidelity scores across all channels ($r_m \geq 0.90$), while the residual gate remains dormant ($\alpha(\boldsymbol{x}) \leq 0.007$). This confirms that on standard uncorrupted inputs, the primary bottleneck representation $\boldsymbol{z}_{\text{base}}$ reliably performs inference without unneeded residual disturbance, mathematically shielding the network from negative transfer.

In Case (3) (illustrated in Figure D.3 top), the speaker utters literal praise (*"I love him as an actor..."*) before concluding with a mocking denial (*"no"*), resulting in an official negative ground truth of $y = -1.00$. Biased by prominent positive lexical tokens, the Base Backbone erroneously predicts a positive sentiment of $\hat{y} = +1.20$ (an error of 2.20). In contrast, CARE-KAN evaluates high physical reliability across both text and acoustic channels ($r_t = 0.89, r_a = 0.87$) while C3D isolates a large directional conflict vector $\boldsymbol{d}_t$. Recognizing genuine communicative incongruity, the gate dynamically opens ($\alpha(\boldsymbol{x}) = 0.084$), routing the conflict residual into the KAN prediction head. Leveraging its localized B-spline activations, KAN accurately navigates the non-smooth decision boundary between literal praise and cynical mockery, outputting $\hat{y} = \mathbf{-0.95}$ (reducing absolute error to merely 0.05).

In Case (4) (illustrated in Figure D.3 bottom), an unperturbed positive review ($y = +3.00$) is subjected to heavy visual Gaussian noise ($\sigma = 1.0$). The Base Backbone, lacking channel-aware iso-

lation, ingests the distorted visual noise and drops its prediction to $+1.25$. Conversely, our router detects high-entropy variance in the visual complement $\boldsymbol{p}_v$ and excessive divergence $\boldsymbol{d}_v$, driving visual reliability down to $r_v \to 0.04$ to trigger the adaptive circuit breaker. Because the uncorrupted textual ($r_t = 0.92$) and acoustic ($r_a = 0.89$) channels remain intact and the gate stays dormant ($\alpha(\boldsymbol{x}) = 0.010$), the primary pathway maintains uncompromised fidelity, outputting a precise prediction of $\hat{y} = +\mathbf{2.92}$.

D.4 Honest Characterization of Failure Modes and Limitations

Despite the empirical and theoretical advantages of CARE-KAN, scientific rigor requires an honest characterization of intrinsic failure modes identified through qualitative inspection:

Failure Mode I: Correlated Multi-Channel Sensor Degradation. CARE-KAN's conflict formulation explicitly relies on a leave-one-out consensus center: $\bar{\boldsymbol{c}}_{-m} = \frac{1}{M-1}\sum_{j \neq m} \boldsymbol{c}_j$. This formulation presumes that at least a subset of sensory channels remains uncompromised to anchor the shared manifold. When environmental adversity triggers simultaneous, highly correlated multi-channel degradation (for example, filming in heavy rain, where acoustic reverberation and visual lens distortion co-occur), the consensus representations of both audio and visual channels diverge erratically in the same directional subspace. In such severe scenarios, the router can experience ambiguity, occasionally misinterpreting the uncorrupted text channel as the conflicting outlier. Future research will explore incorporating Bayesian epistemic priors to anchor unimodal reliability independently of cross-modal consensus.

Failure Mode II: Purely Lexical Sarcasm Without Acoustic/Visual Cues. In conversational communication, speakers occasionally express subtle sarcasm entirely through contextual pragmatics (e.g., deadpan delivery) without exhibiting any measurable shift in vocal pitch or facial micro-expressions. In these instances, the observed physical signals across acoustic and visual streams remain genuinely flat, resulting in near-zero cross-modal conflict vectors ($\boldsymbol{d}_m \approx \mathbf{0}$). Under such conditions, CARE-KAN's gate remains dormant ($\alpha(\boldsymbol{x}) \approx 0$), correctly defaulting to the primary base representation $\boldsymbol{z}_{\text{base}}$. However, this causes the model to inherit the unimodal limitations of the underlying language model, occasionally misclassifying deadpan sarcasm as sincere sentiment. Integrating external conversational discourse history and commonsense knowledge bases into the C3D decomposition represents a promising avenue to address context-dependent affective divergence.