FROM CAPABILITY TO UTILITY: RELIABILITY CONFIGURATION CALIBRATION FOR MULTIMODAL FUSION

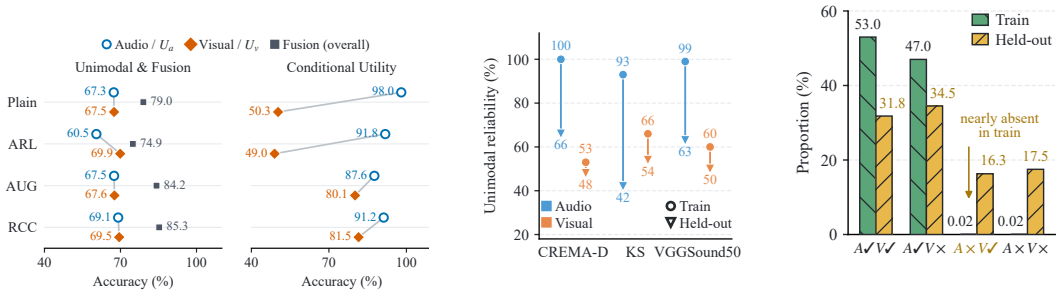
Anonymous authors
Paper under double-blind review

ABSTRACT

Existing research on modality imbalance mainly aims to improve the learning and classification capabilities of individual modalities. However, stronger unimodal capability does not directly ensure that correct evidence is used in fusion. We identify and characterize the capability-to-utility gap (C2U gap): when one modality predicts correctly and the other incorrectly, the fusion model may still make an incorrect prediction. By examining the joint correctness patterns of unimodal predictions, we find a mismatch in reliability configurations between in-sample and held-out data. In particular, critical configurations in which the dominant modality is wrong and the other modality is correct are underrepresented in in-sample predictions. Based on this diagnosis, we propose **R**eliability **C**onfiguration **C**alibration **(RCC)**. RCC uses configuration deficits between in-sample and held-out calibration data to determine the directions and probabilities of feature replacement. Reliability-weighted counterfactual feature replacement provides additional fusion supervision for critical cases, while unimodal anchoring and adoption preservation help maintain unimodal capability and limit interference with previously reliable configurations. Extensive experiments show improved fusion accuracy across six benchmarks and higher visual conditional utility on three audio–visual datasets. Our code is available at `https://anonymous.4open.science/r/RCC`.

1 INTRODUCTION

One of the central goals of multimodal learning is to learn a decision model that effectively combines complementary information from different modalities (Baltrusaitis et al., 2019). This requires both useful predictive representations from each modality and a fusion model that uses them to make correct decisions. However, in audio-visual recognition, we observe that a model may predict correctly using only visual features and incorrectly using only audio features, yet still predict incorrectly when both modalities are fused. To characterize this phenomenon, we measure modality capability by unimodal accuracy, and conditional utility by fusion accuracy on samples where that modality alone predicts correctly while the other modality predicts incorrectly. We call this failure to turn a correct unimodal prediction into a correct fusion decision the **capability-to-utility gap (C2U gap)**.



(a) (b) (c)

Figure 1: **Diagnosing the capability-to-utility gap.** (a) Unimodal and overall fusion accuracy (left), and conditional utility (right), on CREMA-D. (b) Unimodal accuracy on training and held-out samples across CREMA-D, KS, and VGGSound50. (c) Distributions of reliability configurations on training and held-out samples in CREMA-D. Here, U_a and U_v denote fusion accuracy on $A\checkmark V\times$ and $A\times V\checkmark$ samples, respectively; $\checkmark$ and $\times$ indicate correct and incorrect unimodal predictions.

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.

A natural explanation is that the weaker modality is insufficiently learned during joint training (Huang et al., 2022; Xu et al., 2023; Du et al., 2023), a problem widely studied as modality imbalance. Methods such as OGM (Peng et al., 2022), InfoReg (Huang et al., 2025), AUG (Jiang et al., 2025), and MMPareto (Wei & Hu, 2024) improve weak-modality representations and classification through gradient modulation, optimization adjustments, information regularization, or classification capability rebalancing. These methods address whether the weaker modality can make correct predictions on its own. Such improvements, however, do not directly ensure that fusion uses its correct evidence when the dominant modality is wrong. As shown in Fig 1a, methods with nearly identical visual capability can have substantially different visual utility on critical configurations. These observations motivate evaluating conditional utility alongside unimodal capability.

To better understand this gap, we examine reliability configurations, which record the joint correctness of unimodal predictions on the same sample. Fig. 1b shows that in-sample unimodal accuracy exceeds held-out accuracy across the datasets examined, reflecting the optimism of in-sample estimates. Joint correctness statistics in Fig. 1c further show that, for the model examined on CREMA-D, the audio-incorrect, visual-correct configuration is nearly absent from in-sample predictions but occurs much more frequently on held-out data. We call this difference between in-sample and held-out model-induced configuration distributions a **reliability configuration mismatch**. This mismatch can arise from unimodal generalization gaps and captures how correct predictions overlap across modalities.

This diagnosis suggests that limited coverage of critical configurations may weaken the training signal for conditional adoption. Our risk analysis separates changes in configuration frequencies from changes in within-configuration risk differences, identifying conditions under which training risk understates the relative benefit of conditional adoption on held-out samples. Meanwhile, unimodal supervision constrains independent predictions but does not directly constrain fusion behavior when the other modality is present and wrong. This motivates additional fusion supervision for critical configurations to improve the transfer of existing unimodal capability to correct fusion decisions.

Based on this insight, we propose **R**eliability **C**onfiguration **C**alibration **(RCC)**. During calibration, RCC compares the current model's reliability configuration distributions on the training and held-out calibration splits to determine feature replacement directions and probabilities. Using these probabilities, RCC applies targeted counterfactual feature replacement to provide directional fusion supervision, weighted by the predictive reliability of the preserved modality. Unimodal anchoring helps maintain unimodal capability, while adoption preservation adds reverse-direction supervision to counterbalance directional training. RCC thus combines unimodal capability learning with configuration-guided fusion supervision. Our contributions are as follows:

- We distinguish unimodal capability from conditional fusion utility and characterize the C2U gap through reliability configuration mismatch. Our analysis connects configuration coverage to risk comparisons, establishes target-correct enrichment under reliability weighting, and gives sufficient conditions for a positive counterfactual risk advantage.
- We propose RCC, which uses configuration deficits to guide reliability-weighted feature replacement and adds fusion supervision for critical configurations, together with unimodal anchoring and adoption preservation.
- Experiments on multiple multimodal benchmarks show competitive overall performance. Configuration-level diagnostics and ablations further demonstrate improvements in conditional utility and trade-offs between adoption directions.

2 Related Works

Learning with modality imbalance. Differences in optimization and competition between modalities can leave some modalities insufficiently learned during joint training. Existing methods address this problem through optimization, classification capability, information learning, and data organization. G-Blending (Wang et al., 2020) reweights unimodal auxiliary losses and the fusion loss according to their estimated overfitting and generalization behavior, OGM (Peng et al., 2022) dynamically modulates modality gradients, MMPareto (Wei & Hu, 2024) coordinates unimodal and multimodal objectives, and DGL (Wei et al., 2025b) decouples the optimization of modality encoders and the fusion module. PMR (Fan et al., 2023) uses class prototypes to evaluate unimodal performance and guide feature learning in modalities that learn more slowly, AUG (Jiang et al., 2025) improves classification capability through boosting, while ARL (Wei et al., 2025a) uses modality

variances to guide asymmetric optimization. InfoReg (Huang et al., 2025) regulates information acquisition, LFM (Yang et al., 2024) dynamically combines contrastive and supervised learning, IBML (Qin et al., 2026) balances the retention of complementary information, and TCMax (Yu et al., 2026) maximizes total correlation between multimodal features and labels. Data Remixing (Ma et al., 2025) reduces interference between modalities through sample decoupling and batch reassembly. We further distinguish unimodal capability from conditional fusion utility, asking whether existing correct predictive capability translates into a correct fusion decision when the other modality is wrong.

Reliability-aware fusion and calibration. Reliability-aware fusion adjusts evidence integration according to sample quality or predictive uncertainty. TMC (Han et al., 2021) combines multi-view predictions at the evidence level, while PDF (Cao et al., 2024) uses collaborative belief and relative calibration to adjust fusion weights. Probabilistic-circuit fusion (Sidheekh et al., 2025) combines unimodal predictive distributions and estimates modality credibility. CPSC (Jiang et al., 2026) uses conformal prediction to select reliable feature components and calibrates gradients based on instance reliability. At the decision level, Ma & Chen (2025) observe that modality bias can remain even after extensive pretraining of the modality encoders. Burenko & Vetrov (2026) investigate ensembles of independently trained unimodal networks as an alternative to feature fusion. RCC focuses on differences in joint correctness configurations between in-sample and held-out data, using these differences to guide feature replacement during training.

3 METHODS

We propose Reliability Configuration Calibration (RCC) to improve the transfer of unimodal capability to conditional utility by adding fusion supervision for critical reliability configurations. RCC uses configuration deficits estimated during calibration to determine the directions and probabilities of feature replacement. During subsequent training, reliability-weighted feature replacement provides directional supervision, while unimodal anchoring and adoption preservation stabilize training.

3.1 PRELIMINARIES: RELIABILITY CONFIGURATIONS

Given a training set $\mathcal{D} = \{(x_a^i, x_v^i, y_i)\}_{i=1}^N$, where x_a^i and x_v^i are the audio and visual inputs and $y_i \in \{1, \dots, K\}$ is the class label, the modality encoders produce features $a_i = E_a(x_a^i)$ and $v_i = E_v(x_v^i)$. The fusion classifier outputs logits $z_f^i = f_\theta(a_i, v_i)$.

To estimate modality reliability using the same fusion function, we use masked unimodal predictions: $z_a^i = f_\theta(a_i, \mathbf{0})$ and $z_v^i = f_\theta(\mathbf{0}, v_i)$, where $\mathbf{0}$ is a zero feature vector of the corresponding modality dimension. This requires no separate unimodal classifiers and ties reliability estimation to the fusion model used at inference. For each modality $m \in \{a, v\}$, we define its reliability indicator as

$$r_m^i = \mathbf{1}\left[\arg\max_c (z_m^i)_c = y_i\right]. \tag{1}$$

Each sample therefore has a reliability configuration $\mathbf{r}^i = (r_a^i, r_v^i) \in \{0,1\}^2$, corresponding to both modalities being reliable, only audio being reliable, only vision being reliable, or neither being reliable. To measure whether a modality's correct prediction translates into a correct fusion decision when the other modality is wrong, we define audio and visual conditional utility as

$$U_a = \Pr(\hat{y}_f = y \mid r_a = 1, r_v = 0), \qquad U_v = \Pr(\hat{y}_f = y \mid r_a = 0, r_v = 1), \tag{2}$$

where $\hat{y}_f$ denotes the fusion prediction. Modality capability measures predictive performance when using that modality alone, whereas conditional utility measures fusion accuracy when only that modality predicts correctly. A capability-to-utility gap (C2U gap) occurs when one modality predicts correctly and the other incorrectly, yet the fused prediction remains incorrect.

Let $q_{\mathcal{S}}(\mathbf{r})$ denote the empirical distribution of configuration $\mathbf{r}$ on a data split $\mathcal{S}$:

$$q_{\mathcal{S}}(\mathbf{r}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{1}[\mathbf{r}^i = \mathbf{r}]. \tag{3}$$

RCC uses the difference between $q_{\text{train}}(\mathbf{r})$ and $q_{\text{held}}(\mathbf{r})$ to identify reliability configurations that are underrepresented in training. This representation naturally extends to M modalities, with each reliability configuration represented by a binary vector in $\{0,1\}^M$.

3.2 Reliability Configuration Calibration

3.2.1 Deficit Estimation

For audio-visual tasks, the two configurations in which only one modality is reliable, $(0, 1)$ and $(1, 0)$, are most informative about evidence adoption. The configuration $(0, 1)$ means that audio is unreliable while vision is reliable. When this configuration is underrepresented in training, the fusion model receives little supervision on when to adopt visual evidence. We denote the intervention direction for this configuration by $a \to v$: perturbing audio features while preserving vision as the target reliable modality. Similarly, $v \to a$ denotes perturbing vision while preserving audio. For the $a \to v$ direction, we define the configuration deficit as

$$D_{a\to v} = \frac{[q_{\text{held}}(0,1) - q_{\text{train}}(0,1)]_+}{\max(q_{\text{train}}(0,1) + q_{\text{train}}(1,1), \epsilon)}, \tag{4}$$

where $[x]_+ = \max(x, 0)$ and $\epsilon > 0$ prevents division by zero. The denominator is the training accuracy of the target modality, visual, with a minimum value of ϵ. Thus, the deficit measures the configuration shortage relative to the available reliable visual evidence. The deficit $D_{v\to a}$ is defined similarly. The positive part restricts main replacement supervision to configurations that occur less frequently in training than in calibration.

We then map the configuration deficits to replacement probabilities:

$$p_{a\to v} = \min(\alpha D_{a\to v}, p_{\max}), \qquad p_{v\to a} = \min(\alpha D_{v\to a}, p_{\max}), \tag{5}$$

where $\alpha \geq 0$ controls the scale of the replacement probabilities and $p_{\max} \in [0, 1]$ sets their upper limit. Each direction's probability is therefore determined by its own configuration deficit, without specifying in advance which modality needs to be strengthened.

3.2.2 Targeted Counterfactual Replacement

The configuration deficit determines which direction needs additional supervision. We next construct the corresponding training samples. For the $a \to v$ direction, we preserve the visual feature and label of sample i, and replace its audio feature with that of a sample with a different label from the same mini-batch:

$$\tilde{z}^i_{a\to v} = f_\theta(a_{\pi(i)}, v_i), \qquad \pi(i) \neq i, \quad y_{\pi(i)} \neq y_i. \tag{6}$$

The supervision label remains y_i. This construction pairs the original visual feature with audio from another class. The resulting reliability configuration depends on the predictions of both modalities. To emphasize reliable target evidence, we weight replacement supervision by the preserved modality's support for the true class. For modality m, we define the prediction margin δ^i_m of the true class over the other classes and compute a reliability weight C^i_m:

$$\delta^i_m = (z^i_m)_{y_i} - \max_{c\neq y_i}(z^i_m)_c, \qquad C^i_m = \sigma\left(\frac{\delta^i_m - \eta_m}{\tau}\right), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, η_m sets the location of the weight transition, and $\tau > 0$ controls its smoothness. A larger positive margin indicates stronger support for the true class and gives the sample a higher weight. We use continuous weights rather than discarding all samples below a threshold.

For the $a \to v$ direction, we sample $M^i_{a\to v} \sim \text{Bernoulli}(p_{a\to v})$ to indicate whether sample i is selected for replacement in this direction. The loss is

$$\mathcal{L}_{a\to v} = \frac{\sum_i M^i_{a\to v} C^i_v \,\text{CE}(\tilde{z}^i_{a\to v}, y_i)}{\sum_i M^i_{a\to v} C^i_v + \epsilon}, \tag{8}$$

where ϵ is a constant for numerical stability. For the $v \to a$ direction, we similarly preserve audio, replace vision using a sample with a different label, and use C^i_a as the target modality's reliability weight to obtain the normalized loss $\mathcal{L}_{v\to a}$. The total replacement loss is

$$\mathcal{L}_{\text{rcc}} = \mathcal{L}_{a\to v} + \mathcal{L}_{v\to a}. \tag{9}$$

Thus, the configuration deficit determines the direction and sampling probability of replacement supervision, while the preserved modality's reliability determines each selected sample's relative contribution. When computing this additional loss, gradients are stopped through the donor features and reliability weights.

3.2.3 UNIMODAL ANCHORING AND ADOPTION PRESERVATION

Directional replacement teaches the fusion model to use the other modality, but may also affect existing predictive ability. First, the preserved modality must continue to provide discriminative features for replacement supervision to be useful. Second, repeatedly strengthening one direction may make fusion overly favor that modality and impair its performance when the other modality predicts correctly. We address these two effects with unimodal supervision and adoption preservation, respectively.

To maintain each modality's ability to predict independently, we retain unimodal supervision:

$$\mathcal{L}_{\text{uni}} = \mathbb{E}_i \left[\text{CE}(z_a^i, y_i) + \text{CE}(z_v^i, y_i) \right]. \tag{10}$$

This term directly constrains predictions made using only audio or vision, so that directional fusion supervision and unimodal learning proceed together. Adoption preservation (AP) addresses possible shifts in fusion behavior. For example, when the $a \to v$ direction is strengthened, the model more often receives supervision with preserved vision and replaced audio. To maintain its use of audio evidence, we additionally construct samples with preserved audio and replaced vision, supervised with audio reliability weights. The other direction is handled in the same way.

The sampling probabilities for preserving audio and visual adoption are

$$p_{\text{ap}}^a = \rho p_{a \to v}, \qquad p_{\text{ap}}^v = \rho p_{v \to a}, \tag{11}$$

where $\rho \in [0, 1]$ controls the sampling ratio relative to the main replacement direction. The fusion outputs for the two preservation directions are

$$\tilde{z}_{\text{ap}}^{a,i} = f_\theta(a_i, v_{\pi(i)}), \qquad \tilde{z}_{\text{ap}}^{v,i} = f_\theta(a_{\pi(i)}, v_i). \tag{12}$$

The adoption preservation loss is

$$\mathcal{L}_{\text{ap}} = \sum_{m \in \{a,v\}} \frac{\sum_i M_{\text{ap}}^{m,i} C_m^i \, \text{CE}(\tilde{z}_{\text{ap}}^{m,i}, y_i)}{\sum_i M_{\text{ap}}^{m,i} C_m^i + \epsilon}, \tag{13}$$

where $M_{\text{ap}}^{m,i} \sim \text{Bernoulli}(p_{\text{ap}}^m)$. AP follows the main replacement probabilities and can add supervision in the opposite direction even when that direction has no positive configuration deficit. The sampling ratio ρ and loss coefficient λ_{ap} control its contribution to training.

3.2.4 TRAINING OBJECTIVE

RCC combines fusion supervision on the original paired samples, unimodal supervision, directional replacement supervision, and adoption preservation:

$$\mathcal{L} = \mathcal{L}_{\text{fus}} + \beta \mathcal{L}_{\text{uni}} + \lambda_{\text{rcc}} \mathcal{L}_{\text{rcc}} + \lambda_{\text{ap}} \mathcal{L}_{\text{ap}}. \tag{14}$$

The original fusion supervision maintains learning on real paired samples; unimodal supervision preserves each modality's discriminative ability; directional replacement adds fusion supervision for critical configurations; and AP reduces shifts caused by directional training. The coefficients β, λ_{rcc}, and λ_{ap} control the last three terms, respectively.

During calibration, RCC estimates and updates replacement probabilities using predictions from the model being trained. We then fix these probabilities and train the same network architecture from scratch on the full training set. The calibration stage determines replacement directions and probabilities, while reliability weights are computed from the current model's unimodal predictions during training. At inference, only the standard fusion output $f_\theta(a, v)$ is used, without feature replacement or reliability weighting. The two-stage procedure is given in Appendix B.2.

3.3 ANALYSIS OF RELIABILITY CONFIGURATION CALIBRATION

At a fixed training snapshot, we analyze how configuration coverage, reliability weighting, and counterfactual supervision affect the comparison between two candidate fusion rules, h_{dom} and h_{cond}. Let $\Delta_S = R_S(h_{\text{dom}}) - R_S(h_{\text{cond}})$ be their empirical risk difference on split S, with a shared configuration partition, and define $\gamma_S(\mathbf{r})$ analogously within nonempty configurations. For a critical configuration $\mathbf{r}^\star$ present in both splits,

$$\Delta_{\text{held}} - \Delta_{\text{train}} = \big[q_{\text{held}}(\mathbf{r}^\star) - q_{\text{train}}(\mathbf{r}^\star)\big]\gamma_{\text{train}}(\mathbf{r}^\star) + \mathcal{E}. \tag{15}$$

The first term captures the critical configuration's coverage contribution. The remainder accounts for changes within configurations, other configurations, and split-specific support. A positive leading term exceeding $|\mathcal{E}|$ gives $\Delta_{\text{held}} > \Delta_{\text{train}}$, linking a coverage deficit to a larger held-out risk difference under this condition. If the configuration occurs only in held-out data, its contribution is $q_{\text{held}}(\mathbf{r}^\star)\gamma_{\text{held}}(\mathbf{r}^\star)$. Appendix A.1 gives the full decomposition and remainder bound.

For a fixed nonempty set $\mathcal{J}$ of selected replacement samples, let $T_i = \mathbf{1}[\delta_i > 0]$ indicate a correct target prediction, assuming no logit ties. The reliability weight $C_i = \sigma((\delta_i - \eta)/\tau)$, with common η and $\tau > 0$, is increasing in the same margin. Consequently,

$$\widehat{A}_C = \frac{\sum_{i\in\mathcal{J}} C_i T_i}{\sum_{i\in\mathcal{J}} C_i} \geq \frac{1}{|\mathcal{J}|}\sum_{i\in\mathcal{J}} T_i = \widehat{A}. \tag{16}$$

By the nonnegative empirical covariance of C_i and T_i (Lemma A.2), weighting increases or preserves the weight share assigned to target-correct samples. The joint intended configuration additionally depends on the replacement source's prediction, which enters the following risk analysis.

For one replacement direction, let $R_{\text{cf}}(h; p)$ be the expected batch-normalized loss and $\gamma_{\text{cf}}(p)$ its difference between the two rules. Appendix A.3 defines the corresponding effective distribution, retaining finite-batch normalization and zero-loss batches. Let $\bar{\kappa} > 0$ be its mean batch scaling factor and π_{eff} the probability of a correct target and an incorrect replacement source. If the mean loss difference is at least $m > 0$ on this event and at least $-M$, $M \geq 0$, on its complement whenever these events have positive mass, then

$$\gamma_{\text{cf}}(p) \geq \bar{\kappa}\big[(m + M)\pi_{\text{eff}} - M\big]. \tag{17}$$

Under these loss conditions, $\pi_{\text{eff}} > M/(m + M)$ yields a positive counterfactual advantage and strictly increases the directional objective difference when $\lambda_{\text{rcc}} > 0$. Thus, intended-configuration mass and configuration-wise loss differences jointly determine this contribution. The full objective combines it with unimodal supervision, the other replacement direction, and AP. Proofs are provided in Appendix A.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate RCC on six multimodal classification benchmarks: CREMA-D (Cao et al., 2014), Kinetics-Sounds (KS) (Arandjelovic & Zisserman, 2017), VGGSound50 (Chen et al., 2020), and AVSBench (Zhou et al., 2022) for audio–visual classification; Food101 (Wang et al., 2015) for image–text classification; and CMU-MOSI (Zadeh et al., 2016) for audio–visual–text sentiment classification.

Baselines. We compare RCC with representative methods for modality imbalance: OGM_GE (Peng et al., 2022), MMPareto (Wei & Hu, 2024), AUG (Jiang et al., 2025), InfoReg (Huang et al., 2025), ARL (Wei et al., 2025a), DGL (Wei et al., 2025b), CPSC (Jiang et al., 2026), IBML (Qin et al., 2026), and TCMax (Yu et al., 2026).

Evaluation protocols. We report fusion accuracy and unimodal classification accuracy from jointly trained models. For MOSI, we report three-class accuracy and F1 score. Conditional utilities follow Section 3.1, using each model's own reliability configurations unless otherwise specified. The main comparison tables report mean results over three runs.

Implementation details. Following InfoReg (Huang et al., 2025) and IBML (Qin et al., 2026), we use randomly initialized ResNet18 (He et al., 2016) encoders for both modalities in audio–visual tasks and for images in Food101. The Food101 text encoder uses fixed GloVe embeddings and a bidirectional GRU. For MOSI, we apply temporal pooling to the pre-extracted features and use MLP encoders. Training configurations follow prior work. Except for the GloVe embedding layer, all models are trained from scratch using SGD with a momentum of 0.9, a mini-batch size of 64, a learning rate of 1×10^{-3}, and a weight decay of 1×10^{-4}. RCC first holds out 10% of the training set for calibration, then reinitializes the same model architecture and trains it on the full training set.

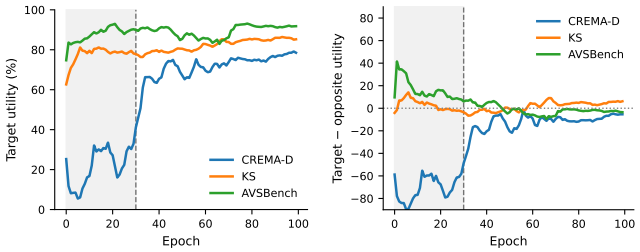
The two stages last 60 and 100 epochs(Appendix B.2), respectively. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. Appendix B provides details of the datasets and input processing.

Table 1: **Results of the average test accuracy(%) of three seeds on five datasets.** Both the results of only using a single modality ("Audio" and "Visual") and the results of combining all modalities ("Multi") are listed. The best results and second best results are **bold** and underlined, respectively. The complete results, including mean ± std, are in Appendix C.

Methods	CREMA-D			Kinetics-Sounds			VGGSound50			AVSBench			Food101		
	Audio	Visual	**Multi**	Audio	Visual	**Multi**	Audio	Visual	**Multi**	Audio	Visual	**Multi**	Visual	Text	**Multi**
OGM-GE (Peng et al., 2022)	53.9	35.1	76.7	49.2	31.6	69.4	40.5	21.4	54.8	74.7	27.6	88.2	81.3	35.4	87.9
MMPareto (Wei & Hu, 2024)	66.9	48.3	73.6	57.0	50.6	73.3	45.0	36.0	58.5	81.0	50.1	90.9	84.0	51.2	90.4
AUG (Jiang et al., 2025)	68.4	67.8	83.3	56.0	**62.4**	76.3	43.6	39.8	58.1	81.3	58.6	91.3	83.7	56.3	90.7
DGL (Wei et al., 2025b)	69.0	53.1	77.1	50.3	59.2	75.3	45.2	39.9	59.9	81.6	**60.8**	92.0	83.6	56.6	90.1
ARL (Wei et al., 2025a)	60.7	**69.0**	77.2	49.4	58.9	75.0	46.2	38.1	60.0	81.4	56.8	91.2	84.3	56.5	89.5
InfoReg (Huang et al., 2025)	68.9	42.8	74.1	56.2	51.3	73.7	45.0	36.8	60.0	81.6	53.0	91.8	84.5	51.8	90.5
CPSC (Jiang et al., 2026)	68.7	60.0	80.3	**58.6**	61.4	74.5	45.1	**42.5**	56.7	79.9	59.3	87.4	83.8	51.2	89.6
IBML (Qin et al., 2026)	65.4	66.0	79.9	48.0	52.0	69.7	43.8	39.8	60.6	79.0	59.3	92.3	83.7	55.5	90.4
TCMax (Yu et al., 2026)	67.5	64.3	79.6	54.7	52.5	76.2	45.8	33.6	60.6	69.0	40.2	87.1	83.6	53.2	90.5
Ours	**69.4**	68.6	**85.0**	57.9	59.4	**76.9**	**47.0**	40.4	**62.0**	**82.0**	59.9	**93.3**	**84.7**	**57.1**	**91.3**

Table 2: **Results of the average test accuracy(%) of three seeds on MOSI datasets.**

Methods	Acc-3	F1
AUG (Jiang et al., 2025)	71.9	60.5
ARL (Wei et al., 2025a)	72.1	59.4
DGL (Wei et al., 2025b)	72.6	61.6
InfoReg (Huang et al., 2025)	70.3	61.5
CPSC (Jiang et al., 2026)	72.0	61.5
IBML (Qin et al., 2026)	72.1	55.3
TCMax (Yu et al., 2026)	72.8	52.6
Ours	**73.4**	**61.9**



(a) (b)

Figure 2: **Evolution of target-direction conditional utility during RCC training.** (a) Conditional utility of the target modality.(b) Signed target–opposite utility gap $G^{\star}$. The shaded area denotes warm-up, and the dashed line marks the activation of RCC.

4.2 Comparison across Different Datasets

Tables 1 and 2 show that RCC achieves leading fusion performance on most datasets, although this advantage is not always accompanied by higher unimodal accuracy. For example, on KS, CPSC achieves higher audio and visual accuracies than RCC but lower fusion accuracy. This shows that stronger classification performance in both branches does not necessarily lead to better fusion results. This observation supports our motivation: beyond improving unimodal learning, we also need to consider how fusion models use existing unimodal capabilities. The overall performance gains of RCC further show that improving fusion training through reliability configuration calibration can benefit final classification performance. Its effectiveness extends from audio-visual tasks to image-text classification on Food101 and trimodal sentiment classification on MOSI, supporting the applicability of this training approach across different modality combinations.

Table 3: **Conditional utility (%) on three datasets.** For each method, reliability configurations are induced by its own unimodal predictions, so the resulting evaluation subsets may differ across methods. U_a and U_v denote fusion accuracy on the corresponding $A\checkmark V\times$ and $A\times V\checkmark$ subsets, respectively; $\mathrm{Asym} = |U_a - U_v|$. Pairwise evaluations of RCC and the selected baselines on identical configuration intersections in CREMA-D are reported in Appendix E.

Models	CREMA-D			KS			AVSBench		
	$\mathbf{U_a}$	$\mathbf{U_v}$	**Asym** ↓	$\mathbf{U_a}$	$\mathbf{U_v}$	**Asym** ↓	$\mathbf{U_a}$	$\mathbf{U_v}$	**Asym** ↓
AUG (Jiang et al., 2025)	87.42	74.48	12.94	88.21	84.45	3.76	93.95	91.66	2.29
ARL (Wei et al., 2025a)	91.80	48.96	42.84	88.81	80.30	8.51	92.07	87.95	4.12
DGL (Wei et al., 2025b)	94.47	38.77	55.70	91.19	75.77	15.42	95.62	85.48	10.14
IBML (Qin et al., 2026)	93.54	65.41	28.13	89.23	74.09	15.14	96.34	82.27	14.07
Ours	85.71	81.88	**3.83**	86.08	87.15	**1.07**	93.82	93.81	**0.01**

4.3 Adoption Utility Analysis

Table 3 reports conditional utility on the reliability configurations defined by each model's own unimodal predictions. The resulting model-level profiles show that RCC has the highest observed U_v and the lowest utility asymmetry among the reported methods. Compared with AUG, RCC has higher visual conditional utility but lower U_a on CREMA-D and KS, while its U_a remains comparable on AVSBench. Because the conditional subsets are defined separately by each model, these results provide model-level diagnostics and do not by themselves establish differences in fusion performance on identical samples. To further compare performance on the same samples, Appendix E reports pairwise evaluations of RCC and selected baselines on common configurations in CREMA-D. Each model pair is evaluated on the intersections of their corresponding configuration subsets, and the results likewise show gains in visual utility accompanied by losses in audio utility.

Fig. 2 tracks target-direction conditional utility and the signed utility gap during RCC training. Here, $G^\star = U_v - U_a$ for $a \to v$ and $G^\star = U_a - U_v$ for $v \to a$; a smaller absolute gap indicates closer utilities in the two directions. After RCC is activated, CREMA-D shows the clearest increase in target utility and a substantial reduction in its initial directional gap, while KS and AVSBench maintain high target utility with relatively small gaps.

Table 4: **Reliability-configuration mismatch across datasets.** $A \times V\checkmark$ denotes audio-wrong visual-correct samples, and $A\checkmark V\times$ denotes audio-correct visual-wrong samples. PG (Positive gap) denotes $[q_{\text{held}}(\mathbf{r}) - q_{\text{train}}(\mathbf{r})]_+$.

Dataset	Train (%) $A \times V\checkmark$	Heldout (%) $A \times V\checkmark$	PG (%) $A \times V\checkmark$	p_{a2v} (%)	Train (%) $A\checkmark V\times$	Heldout (%) $A\checkmark V\times$	PG (%) $A\checkmark V\times$	p_{v2a} (%)	Dominant direction
CREMA-D	0.00	15.22	15.22	10.07	47.06	31.94	0.00	0.00	$A \times V\checkmark$ / a2v
KS	2.98	27.18	24.20	5.89	36.47	19.84	0.00	0.00	$A \times V\checkmark$ / a2v
AVSBench	0.00	7.25	7.25	1.52	28.64	42.03	13.38	2.01	$A\checkmark V\times$ / v2a
VGGSound50	0.49	16.38	15.89	1.16	31.39	25.13	0.00	0.00	$A \times V\checkmark$ / a2v

4.4 Reliability Configuration Diagnosis and Calibration

Table 4 compares reliability configuration distributions on training and held-out calibration samples, together with the resulting replacement probabilities. On CREMA-D, KS, and VGGSound50, $A\times V\checkmark$ occurs less frequently in in-sample predictions than on held-out calibration samples, while there is no positive gap in the opposite direction. RCC therefore assigns a nonzero replacement probability to $a \to v$, providing more fusion supervision for cases in which only the visual prediction is correct.

AVSBench has positive gaps in both configurations where only one modality predicts correctly. Both directions therefore receive nonzero replacement probabilities, with a higher probability for $v \to a$. These results show that RCC can assign replacements to one or both directions based on the observed configuration differences, providing targeted fusion supervision for critical cases that are underrepresented during training. Complete reliability configuration distributions show in Appendix D.

Table 5: **Ablations of loss terms and reliability weighting in RCC.** *w/o* removes the indicated loss; *Unweighted* sets its sample reliability weights to one without changing the loss coefficients. Other components are retained.

Variant	**CREMA-D**					**KS**				
	Audio	Visual	Multi	$\mathbf{U_a}$	$\mathbf{U_v}$	Audio	Visual	Multi	$\mathbf{U_a}$	$\mathbf{U_v}$
w/o $\mathcal{L}_{\text{rcc}}$	67.91	67.21	79.12	**95.53**	48.36	55.52	57.24	75.14	**89.13**	79.38
w/o $\mathcal{L}_{\text{uni}}$	66.71	30.78	73.39	89.39	62.71	52.56	39.48	66.48	84.62	82.18
w/o $\mathcal{L}_{\text{ap}}$	68.06	67.02	82.64	83.71	80.85	56.33	59.25	76.18	83.22	85.43
Unweighted $\mathcal{L}_{\text{rcc}}$	69.13	67.56	83.20	83.15	80.65	55.26	58.60	76.40	84.11	86.55
Unweighted $\mathcal{L}_{\text{ap}}$	68.86	68.01	83.29	84.17	80.20	55.50	58.32	76.57	86.01	86.88
Unweighted $\mathcal{L}_{\text{rcc}}, \mathcal{L}_{\text{ap}}$	67.27	65.88	82.29	83.99	77.52	54.11	56.95	74.14	82.21	84.26
RCC (Ours)	**69.35**	**68.55**	**84.95**	85.71	**81.88**	**57.85**	**59.42**	**76.85**	86.08	**87.15**

4.5 Ablation Study

Table 5 examines the contributions of individual loss terms and reliability weighting. Removing $\mathcal{L}_{\text{rcc}}$ reduces visual conditional utility and fusion accuracy on both datasets while increasing audio conditional utility, supporting the role of the main replacement supervision in improving adoption when only vision is correct. Removing $\mathcal{L}_{\text{uni}}$ leads to the largest drops in visual and fusion accuracy, highlighting the importance of unimodal anchoring in maintaining predictive capability. Removing $\mathcal{L}_{\text{ap}}$ also lowers fusion accuracy and both conditional utilities, indicating that adoption preservation helps maintain fusion performance in both configurations. Removing reliability weighting from either replacement loss or both reduces fusion accuracy and both conditional utilities, supporting the use of the preserved modality's reliability to weight replacement supervision. Full RCC achieves the highest fusion accuracy among the listed variants.

Table 6: **Replacement strategy controls.** Overall denotes test fusion accuracy (%). $\widetilde{U}_a$ and $\widetilde{U}_v$ are fusion accuracies on fixed Plain-defined $A\checkmark V\times$ and $A\times V\checkmark$ subsets.

	CREMA-D			**AVSBench**		
Strategy	Overall	$\widetilde{U}_a$	$\widetilde{U}_v$	Overall	$\widetilde{U}_a$	$\widetilde{U}_v$
(A) Calibrated versus validation-selected replacement						
Plain	77.69	92.76	51.37	91.08	88.35	95.46
Fixed $a\to v$	81.45	84.21	76.03	90.41	85.54	**97.73**
Fixed $v\to a$	79.84	**94.08**	57.53	91.89	**89.96**	94.32
Symmetric ($0.5\times$)	81.72	87.50	70.55	90.68	86.75	96.59
Calibrated RCC	**82.86**	81.58	**77.40**	**92.81**	88.15	97.46
(B) Exact-K replacement-event controls						
Guided	**81.99**	79.61	**76.71**	**92.81**	86.75	**95.46**
Symmetric	81.05	88.16	68.49	90.68	87.15	93.18
Reverse	79.97	**93.42**	57.53	92.16	**89.96**	94.32

4.6 Configuration-Guided Replacement

Table 6 compares calibrated RCC with (A) validation-selected fixed strategies and (B) event-budget-matched allocations. In Panel A, Symmetric uses a $0.5\times$ loss-coefficient multiplier, while the other replacement strategies use $1.0\times$. Panel B matches replacement-event counts, active batches, AP sampling rules, and loss coefficients, with $K = 7$ for CREMA-D and $K = 2$ for AVSBench. Calibrated RCC exceeds the strongest fixed strategy by 1.14% and 0.92% points on CREMA-D and AVSBench, respectively. Guided allocation also achieves the highest overall accuracy in Panel B, with higher $\widetilde{U}_v$ and lower $\widetilde{U}_a$ than symmetric and reversed allocation. These results support configuration-guided replacement, with fusion gains that persist when replacement-event counts are controlled. Appendix J provides complementary comparisons under the original training protocol, including random replacement without class constraints or reliability weighting.

Additional analyses. Appendices E–I provide additional analyses of common configuration subsets, fusion heads, global modality reweighting, and fusion learning with frozen encoders. We also examine reliability-weighted replacement samples and computational costs to further characterize RCC's fusion behavior and supervision design.

5 Conclusion

We propose RCC to address the capability-to-utility gap in multimodal fusion. RCC uses differences in reliability configuration distributions between in-sample and held-out data to guide feature replacement, and combines reliability-weighted fusion supervision with unimodal anchoring and adoption preservation. Experiments across six benchmarks demonstrate improved fusion performance, while configuration-level analyses show higher visual conditional utility on the three audio–visual benchmarks examined. RCC introduces an additional calibration stage and uses fixed replacement probabilities during subsequent training. Future work will explore more efficient configuration estimation and adaptive replacement strategies to reduce training cost and respond to changes in modality reliability.

AI USE STATEMENT

We used generative AI tools to assist with translation, method implementation, feedback on research methodology and experimental design, interpretation of experimental results, and the formulation and revision of theoretical claims and proofs. We also used these tools for literature search. We reviewed all AI-assisted work and take responsibility for the final content of this study, including any text, claims generated with the help of generative AI.

REPRODUCIBILITY STATEMENT

All datasets used in our experiments are publicly available. We will release our source code, including the model framework and checkpoints. Appendix B.1–B.2 describes input preprocessing and the two-stage training procedure. Appendix B.3 provides hyperparameter settings.

REFERENCES

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019.

Ilya Burenko and Dmitry Vetrov. Beyond modality fusion: Deep ensembles for multimodal classification. *arXiv preprint arXiv:2607.05019*, 2026.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *Forty-first International Conference on Machine Learning, ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5608–5628, 2024.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning*, pp. 8632–8656, 2023.

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.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *9th International Conference on Learning Representations*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Chengxiang Huang, Yake Wei, Zequn Yang, 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. 25854–25863, 2025.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pp. 9226–9259, 2022.

Qing-Yuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. In *Advances in Neural Information Processing Systems*, 2025.

Xun Jiang, Yufan Gu, Disen Hu, Yuqing Hou, Yazhou Yao, Fumin Shen, Heng Tao Shen, and Xing Xu. Multimodal learning on low-quality data with conformal predictive self-calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Xiaoyu Ma and Hao Chen. Revisit modality imbalance at the decision layer. *CoRR*, 2025.

Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. In *Forty-second International Conference on Machine Learning, ICML*, 2025.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

Yang Qin, Yanglin Feng, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Deep information-balanced multimodal learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(8):9384–9396, 2026.

Sahil Sidheekh, Pranuthi Tenali, Saurabh Mathur, Erik Blasch, Kristian Kersting, and Sriraam Natarajan. Credibility-aware multimodal fusion using probabilistic circuits. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pp. 2305–2313, 2025.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12692–12702, 2020.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia & Expo Workshops*, pp. 1–6, 2015.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2250–2259, 2025a.

Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22879–22888, 2025b.

Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *Forty-first International Conference on Machine Learning*, 2024.

Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *Advances in Neural Information Processing Systems*, 2024.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. In *International Conference on Learning Representations*, volume 2026, pp. 45101–45117, 2026.

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.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pp. 386–403, 2022.

A Theoretical Analysis and Proofs

A.1 Configuration Coverage and Train–Held-Out Risk Differences

Common setup. Throughout this appendix, we compare candidate fusion rules h_{dom} and h_{cond} at a fixed training snapshot. Its modality features, reliability weights, and distribution of replacement construction are shared between the rules. The configuration function $\mathbf{r}(z, y) \in \mathcal{R} = \{0, 1\}^2$, defined by the snapshot's unimodal predictions, is also shared across rules and data splits. The rules predict from modality features, while configurations group the samples for analysis. Let ℓ be a nonnegative prediction loss, and assume that all risks defined below are finite.

For each nonempty split $S \in \{\mathrm{train}, \mathrm{held}\}$, let $I_S(\mathbf{r})$ be the indices with configuration $\mathbf{r}$, and let $q_S(\mathbf{r}) = |I_S(\mathbf{r})|/|S|$. Define

$$R_S(h) = \frac{1}{|S|} \sum_{i \in S} \ell(h(z_i), y_i). \tag{18}$$

Only when $q_S(\mathbf{r}) > 0$, define

$$R_S(h \mid \mathbf{r}) = \frac{1}{|I_S(\mathbf{r})|} \sum_{i \in I_S(\mathbf{r})} \ell(h(z_i), y_i), \qquad \gamma_S(\mathbf{r}) = R_S(h_{\mathrm{dom}} \mid \mathbf{r}) - R_S(h_{\mathrm{cond}} \mid \mathbf{r}). \tag{19}$$

Their empirical risk difference is

$$\Delta_S = R_S(h_{\mathrm{dom}}) - R_S(h_{\mathrm{cond}}) = \sum_{\mathbf{r}: q_S(\mathbf{r}) > 0} q_S(\mathbf{r}) \gamma_S(\mathbf{r}). \tag{20}$$

The frequencies and risks in this subsection describe the observed splits; an empty configuration refers to zero observed samples.

Proposition A.1 (Configuration decomposition of the risk-difference shift)**.** *Define the common support and the two split-specific supports by*

$$\begin{aligned} \mathcal{R}_\cap &= \{\mathbf{r} : q_{\mathrm{train}}(\mathbf{r}) > 0,\ q_{\mathrm{held}}(\mathbf{r}) > 0\}, \\ \mathcal{R}_{\mathrm{held}} &= \{\mathbf{r} : q_{\mathrm{train}}(\mathbf{r}) = 0,\ q_{\mathrm{held}}(\mathbf{r}) > 0\}, \\ \mathcal{R}_{\mathrm{train}} &= \{\mathbf{r} : q_{\mathrm{train}}(\mathbf{r}) > 0,\ q_{\mathrm{held}}(\mathbf{r}) = 0\}. \end{aligned} \tag{21}$$

Then the exact decomposition is

$$\begin{aligned} \Delta_{\mathrm{held}} - \Delta_{\mathrm{train}} = &\sum_{\mathbf{r} \in \mathcal{R}_\cap} [q_{\mathrm{held}}(\mathbf{r}) - q_{\mathrm{train}}(\mathbf{r})] \gamma_{\mathrm{train}}(\mathbf{r}) \\ &+ \sum_{\mathbf{r} \in \mathcal{R}_\cap} q_{\mathrm{held}}(\mathbf{r}) [\gamma_{\mathrm{held}}(\mathbf{r}) - \gamma_{\mathrm{train}}(\mathbf{r})] \\ &+ \sum_{\mathbf{r} \in \mathcal{R}_{\mathrm{held}}} q_{\mathrm{held}}(\mathbf{r}) \gamma_{\mathrm{held}}(\mathbf{r}) - \sum_{\mathbf{r} \in \mathcal{R}_{\mathrm{train}}} q_{\mathrm{train}}(\mathbf{r}) \gamma_{\mathrm{train}}(\mathbf{r}). \end{aligned} \tag{22}$$

Every conditional risk in this identity is evaluated on a nonempty configuration.

Proof. Expand both risk differences using Eq. (20). On the common support, add and subtract $q_{\mathrm{held}}(\mathbf{r}) \gamma_{\mathrm{train}}(\mathbf{r})$ to obtain

$$\begin{aligned} &q_{\mathrm{held}}(\mathbf{r}) \gamma_{\mathrm{held}}(\mathbf{r}) - q_{\mathrm{train}}(\mathbf{r}) \gamma_{\mathrm{train}}(\mathbf{r}) \\ &\quad = [q_{\mathrm{held}}(\mathbf{r}) - q_{\mathrm{train}}(\mathbf{r})] \gamma_{\mathrm{train}}(\mathbf{r}) + q_{\mathrm{held}}(\mathbf{r}) [\gamma_{\mathrm{held}}(\mathbf{r}) - \gamma_{\mathrm{train}}(\mathbf{r})]. \end{aligned} \tag{23}$$

Configurations present in only one split retain their original contribution, with a positive sign for held-out and a negative sign for training. Summing gives Eq. (22). □

A critical configuration and the remainder. For $\mathbf{r}^\star \in \mathcal{R}_\cap$, isolate its coverage contribution:

$$\Delta_{\mathrm{held}} - \Delta_{\mathrm{train}} = A_\star + \mathcal{E}, \qquad A_\star = [q_{\mathrm{held}}(\mathbf{r}^\star) - q_{\mathrm{train}}(\mathbf{r}^\star)] \gamma_{\mathrm{train}}(\mathbf{r}^\star). \tag{24}$$

Here $\mathcal{E}$ consists of all remaining terms of Eq. (22), including the within-configuration change at $\mathbf{r}^\star$. By the triangle inequality,

$$\begin{aligned}|\mathcal{E}| \le & \sum_{\mathbf{r}\in\mathcal{R}_\cap\setminus\{\mathbf{r}^\star\}} |q_{\text{held}}(\mathbf{r}) - q_{\text{train}}(\mathbf{r})|\,|\gamma_{\text{train}}(\mathbf{r})| \\ & + \sum_{\mathbf{r}\in\mathcal{R}_\cap} q_{\text{held}}(\mathbf{r})|\gamma_{\text{held}}(\mathbf{r}) - \gamma_{\text{train}}(\mathbf{r})| \\ & + \sum_{\mathbf{r}\in\mathcal{R}_{\text{held}}} q_{\text{held}}(\mathbf{r})|\gamma_{\text{held}}(\mathbf{r})| + \sum_{\mathbf{r}\in\mathcal{R}_{\text{train}}} q_{\text{train}}(\mathbf{r})|\gamma_{\text{train}}(\mathbf{r})| =: B_{\mathcal{E}}.\end{aligned} \tag{25}$$

Thus, $A_\star > B_{\mathcal{E}}$ is sufficient for $\Delta_{\text{held}} > \Delta_{\text{train}}$. Combined with $\Delta_{\text{train}} \ge 0$, it also establishes $\Delta_{\text{held}} > 0$. Approximate stability of configuration-wise risk differences makes the associated remainder term small; that term remains part of the exact decomposition.

If $\mathbf{r}^\star \in \mathcal{R}_{\text{held}}$, the valid identity is instead

$$\Delta_{\text{held}} - \Delta_{\text{train}} = q_{\text{held}}(\mathbf{r}^\star)\gamma_{\text{held}}(\mathbf{r}^\star) + \mathcal{E}', \tag{26}$$

where $\mathcal{E}'$ contains all other contributions. The training risk has no direct term for this configuration. This covers complete empirical absence; a small nonzero training frequency belongs to the common-support case.

Relation to the coverage-only argument. If the two rules have equal conditional training risks outside $\mathbf{r}^\star$, then $\Delta_{\text{train}} = q_{\text{train}}(\mathbf{r}^\star)\gamma_{\text{train}}(\mathbf{r}^\star)$ when the configuration is present, and $\Delta_{\text{train}} = 0$ when it is absent. A small frequency gives a small risk contribution when the conditional loss difference is controlled. For cross-entropy, one sufficient condition is a lower bound $c > 0$ on both rules' true-class probabilities, which gives losses in $[0, -\log c]$. The general decomposition additionally retains changes within and outside the critical configuration, separating their contributions from frequency changes.

A.2 Reliability Weighting and Target-Correct Enrichment

For one replacement direction, fix a nonempty selected set $\mathcal{J}$ after donor construction and sampling, and let $n = |\mathcal{J}|$. Use the preserved target modality's true-class margins from the common snapshot. Assuming no ties between the true class and a competing maximum, $T_i = \mathbf{1}[\delta_i > 0]$ is its correctness indicator.

Lemma A.2 (Target-correct enrichment within a selected set)**.** *Let $C_i = C(\delta_i) \ge 0$ be a common nondecreasing function of the margin, with $\sum_{i\in\mathcal{J}} C_i > 0$. In particular, Eq. (7) satisfies this condition for common η and $\tau > 0$. Define*

$$\widehat{A} = \frac{1}{n}\sum_{i\in\mathcal{J}} T_i, \qquad \widehat{A}_C = \frac{\sum_{i\in\mathcal{J}} C_i T_i}{\sum_{i\in\mathcal{J}} C_i}, \qquad \overline{C} = \frac{1}{n}\sum_{i\in\mathcal{J}} C_i. \tag{27}$$

With empirical covariance defined using divisor n, we have

$$\widehat{A}_C - \widehat{A} = \frac{\mathrm{Cov}_n(C,T)}{\overline{C}} \ge 0. \tag{28}$$

Proof. By expanding the covariance,

$$\mathrm{Cov}_n(C,T) = \frac{1}{n}\sum_i C_i T_i - \overline{C}\,\widehat{A} = \frac{1}{2n^2}\sum_{i,j\in\mathcal{J}} (C_i - C_j)(T_i - T_j). \tag{29}$$

Both $C(\delta)$ and $\mathbf{1}[\delta > 0]$ are nondecreasing functions of the same scalar. Every product in the final sum is therefore nonnegative. Dividing by $\overline{C} > 0$ proves the result. □

Interpretation. The inequality holds for each realized selected set and describes the target-correct share of its normalized sample weights. The denominator $\sum_i C_i + \epsilon$ rescales all weighted terms equally, preserving their relative weight shares. Batches discarded for negligible total weight are handled in the next subsection.

The intended configuration requires both a correct target prediction and a replacement source that predicts the recipient label incorrectly. The next result therefore considers the joint-configuration probability together with the loss differences between the candidate fusion rules.

A.3 A Sufficient Condition for Counterfactual Risk Advantage

Actual batch normalization. Consider $a \to v$; the other direction is symmetric. Let ω collect mini-batch sampling, valid cross-class donor selection, and Bernoulli masks under the common setup. Write $\tilde{z}_i = (a_{\pi(i)}, v_i)$ and retain label y_i, where $\pi(i)$ is the donor map. Set

$$W_\omega = \sum_i M_i C_i, \qquad \kappa_\omega = \mathbf{1}[W_\omega > \varepsilon_0] \frac{W_\omega}{W_\omega + \epsilon}, \qquad \epsilon > 0, \quad \varepsilon_0 \geq 0. \tag{30}$$

The ideal loss in Eq. (8) corresponds to $\varepsilon_0 = 0$. The implementation returns zero for an empty selection or total selected weight $W_\omega \leq 10^{-8}$ and otherwise uses denominator $W_\omega + 10^{-8}$; this is represented by $\varepsilon_0 = \epsilon = 10^{-8}$. This safeguard acts on the batch's total selected weight, while individual reliability weights remain continuous.

For an active batch, define $Q_\omega(i) = M_i C_i / W_\omega$. On inactive batches, expressions multiplied by $\kappa_\omega = 0$ are defined to be zero. The risk is

$$\begin{aligned} R_{\mathrm{cf}}(h; p) &= \mathbb{E}_\omega \left[\mathbf{1}[W_\omega > \varepsilon_0] \frac{\sum_i M_i C_i \ell(h(\tilde{z}_i), y_i)}{W_\omega + \epsilon} \right] \\ &= \mathbb{E}_\omega \left[\kappa_\omega \sum_i Q_\omega(i) \ell(h(\tilde{z}_i), y_i) \right]. \end{aligned} \tag{31}$$

Define the sample-wise and expected loss differences by

$$\begin{aligned} d_i &= \ell(h_{\mathrm{dom}}(\tilde{z}_i), y_i) - \ell(h_{\mathrm{cond}}(\tilde{z}_i), y_i), \\ \gamma_{\mathrm{cf}}(p) &= R_{\mathrm{cf}}(h_{\mathrm{dom}}; p) - R_{\mathrm{cf}}(h_{\mathrm{cond}}; p). \end{aligned} \tag{32}$$

If $\bar{\kappa} = \mathbb{E}_\omega[\kappa_\omega] = 0$, both counterfactual risks and their difference are zero. For the remaining case $\bar{\kappa} > 0$, define the effective distribution on batch–sample pairs by

$$\mathbb{E}_{\widetilde{Q}}[g] = \frac{\mathbb{E}_\omega[\kappa_\omega \sum_i Q_\omega(i) g(\omega, i)]}{\bar{\kappa}}. \tag{33}$$

This is a probability distribution because the active Q_ω sums to one. Under the common risk assumptions, d is integrable under $\widetilde{Q}$, so linearity gives

$$\gamma_{\mathrm{cf}}(p) = \mathbb{E}_\omega[\kappa_\omega \sum_i Q_\omega(i) d_i] = \bar{\kappa}\, \mathbb{E}_{\widetilde{Q}}[d]. \tag{34}$$

This representation retains the batchwise normalization; pooling weights into the global ratio $\mathbb{E}[MCd]/\mathbb{E}[MC]$ generally defines a different quantity.

Proposition A.3 (Positive counterfactual advantage under configuration-wise loss conditions)**.** *Under the common setup with $\bar{\kappa} > 0$, let $E^\star$ be the event that the preserved target predicts the recipient label correctly and the replacement source predicts that label incorrectly. Set $\pi_{\mathrm{eff}} = \Pr_{\widetilde{Q}}(E^\star)$. If constants $m > 0$ and $M \geq 0$ satisfy*

$$\mathbb{E}_{\widetilde{Q}}[d \mathbf{1}_{E^\star}] \geq m \pi_{\mathrm{eff}}, \qquad \mathbb{E}_{\widetilde{Q}}[d \mathbf{1}_{(E^\star)^c}] \geq -M(1 - \pi_{\mathrm{eff}}), \tag{35}$$

then

$$\gamma_{\mathrm{cf}}(p) \geq \bar{\kappa}[(m + M)\pi_{\mathrm{eff}} - M]. \tag{36}$$

In particular, $\pi_{\mathrm{eff}} > M/(m + M)$ is sufficient for $\gamma_{\mathrm{cf}}(p) > 0$.

Proof. Partition the integrable loss difference by $E^\star$ and its complement. Equation (35) gives

$$\mathbb{E}_{\widetilde{Q}}[d] = \mathbb{E}_{\widetilde{Q}}[d \mathbf{1}_{E^\star}] + \mathbb{E}_{\widetilde{Q}}[d \mathbf{1}_{(E^\star)^c}] \geq m \pi_{\mathrm{eff}} - M(1 - \pi_{\mathrm{eff}}). \tag{37}$$

Multiply by $\bar{\kappa} > 0$ and use Eq. (34). Since $m + M > 0$, the displayed threshold makes the lower bound strictly positive. □

Conditional means and boundary cases. For $0 < \pi_{\mathrm{eff}} < 1$, the moment conditions are equivalent to $\mathbb{E}_{\widetilde{Q}}[d \mid E^\star] \geq m > 0$ and $\mathbb{E}_{\widetilde{Q}}[d \mid (E^\star)^c] \geq -M$. They constrain configuration-wise means. At $\pi_{\mathrm{eff}} = 1$, the complementary event has zero mass, its moment condition is vacuous, and the bound becomes $\gamma_{\mathrm{cf}} \geq \bar{\kappa} m > 0$. At $\pi_{\mathrm{eff}} = 0$, the target event condition is vacuous and the bound becomes $-\bar{\kappa} M$, which gives no positive certificate. The moment formulation covers both boundaries without conditioning on null events.

Connecting construction diagnostics to the bound. Let A_Q and B_Q be the probabilities of target correctness and replacement-source incorrectness under the same weighted distribution Q. Then

$$\Pr_Q(E^\star) \geq \max(0, A_Q + B_Q - 1), \tag{38}$$

by the union bound on the complementary events. Joint per-sample records allow direct measurement of $\Pr_Q(E^\star)$. When $Q = \widetilde{Q}$, this bounds π_{eff}. A pooled diagnostic with a different normalization instead describes its own joint proportion π_{diag}; interpreting it as π_{eff} requires matching the weighting distribution. Together, the target marginal in Lemma A.2 and the joint statistic describe two aspects of construction quality. Applying the risk-advantage condition additionally requires the configuration-wise loss quantities m and M.

A.4 The Full Objective and Sampling Probabilities

Corollary A.4 (Contribution to the objective difference)**.** *For $\lambda_{\text{rcc}} \geq 0$, define the one-direction augmented objective*

$$R_+(h) = R_{\text{train}}(h) + \lambda_{\text{rcc}} R_{\text{cf}}(h; p). \tag{39}$$

Then

$$\Delta R_+ := R_+(h_{\text{dom}}) - R_+(h_{\text{cond}}) = \Delta_{\text{train}} + \lambda_{\text{rcc}} \gamma_{\text{cf}}(p). \tag{40}$$

A positive $\gamma_{\text{cf}}(p)$ and $\lambda_{\text{rcc}} > 0$ strictly increase this risk difference relative to Δ_{train}. They ensure $\Delta R_+ > 0$ if additionally $\Delta_{\text{train}} \geq 0$, or more generally if $\lambda_{\text{rcc}} \gamma_{\text{cf}}(p) > -\Delta_{\text{train}}$.

Proof. Subtract the two definitions of R_+ and collect the empirical and counterfactual risk differences. The inequalities follow directly. □

Combining the training terms. Let R_{other} collect unimodal supervision, counterfactual supervision in the other direction, and AP, with their actual loss coefficients. Then

$$\begin{aligned} R_{\text{full}}(h) &= R_{\text{train}}(h) + \lambda_{\text{rcc}} R_{\text{cf}}(h; p) + R_{\text{other}}(h), \\ \Delta R_{\text{full}} &= \Delta_{\text{train}} + \lambda_{\text{rcc}} \gamma_{\text{cf}}(p) + \Delta R_{\text{other}}, \end{aligned} \tag{41}$$

where $\Delta R_{\text{other}} = R_{\text{other}}(h_{\text{dom}}) - R_{\text{other}}(h_{\text{cond}})$. The full objective favors h_{cond} exactly when the sum of these contributions is positive.

Sampling frequency and normalization. Condition on a batch with K candidates having valid donors and sample independent masks with a common probability p. Then

$$\mathbb{E}[N_{\text{selected}} \mid K] = Kp, \qquad \Pr(N_{\text{selected}} > 0 \mid K) = 1 - (1-p)^K. \tag{42}$$

For $K = 0$, both quantities are zero. Both are nondecreasing in a nonnegative deficit when $p = \min(\alpha D, p_{\max})$ with $\alpha \geq 0$ and $0 \leq p_{\max} \leq 1$. The second expression describes nonempty selection; the numerical weight safeguard further determines whether the loss is active. The effect on normalized loss can be seen from a simple example: on a fixed batch of $B > 0$ candidates with unit weights $C_i = 1$ and fixed losses ℓ_i, the ideal version with $\varepsilon_0 = 0$ satisfies

$$\mathbb{E}[\mathcal{L} \mid \ell_1, \ldots, \ell_B] = \mathbb{E}\left[\frac{N_{\text{selected}}}{N_{\text{selected}} + \epsilon}\right] \frac{1}{B} \sum_{i=1}^{B} \ell_i. \tag{43}$$

Conditional on a positive selected count, all subsets of that size are equally likely, which gives this identity. As $\epsilon \to 0$, the factor becomes $1 - (1-p)^B$. With unequal weights, sampling also affects the effective distribution in Eq. (33). The two RCC directions are normalized separately and then summed, as are the AP directions. Equal sums of sampling probabilities therefore do not imply equal effective auxiliary supervision or equal gradients. The coefficient λ_{rcc} is the explicit replacement-loss multiplier, whereas p controls selection and the induced normalized objective.

The results provide conditional risk comparisons at a fixed training snapshot, explaining the roles of configuration coverage, reliability weighting, and counterfactual supervision. The deficit normalization and probability mapping in Eqs. (4)–(5) are design choices; the analysis does not establish their optimality or identify a causal explanation for the observed C2U gap.

B Data and Experimental Details

B.1 Datasets and Input Processing

We evaluate RCC on the following six datasets. Calibration samples are drawn only from the original training set; test samples are not used to estimate the RCC calibration probabilities. All compared methods are retrained and evaluated using the same dataset versions and splits as RCC.

CREMA-D. CREMA-D (Cao et al., 2014) is an audio-visual emotion recognition dataset containing 7,442 short videos recorded by 91 actors. It covers six emotions: anger, disgust, fear, happiness, neutrality, and sadness. We use the same data split as OGM (Peng et al., 2022). Audio is resampled to 22,050 Hz and repeated or truncated to form a 20 s input. The visual input is one frame read from the cached video frames.

Kinetics-Sounds (KS). KS (Arandjelovic & Zisserman, 2017) is an audio-visual action recognition dataset constructed by selecting action classes with related audio and visual content from Kinetics (Kay et al., 2017). Its video clips are collected from YouTube and are approximately 10 s long. We use the same data split as OGM (Peng et al., 2022). Audio is sampled at 16,000 Hz. Clips shorter than 10 s are repeated to reach at least 10 s, and a random 5 s segment is then used as input. The visual modality uses three frames. Random audio cropping is applied during both training and testing.

AVSBench. AVSBench (Zhou et al., 2022) is an audio-visual segmentation dataset with single-source and multi-source subsets. The single-source subset, S4, contains more than 4,000 videos across 23 classes. We use its clip-level class labels for audio-visual classification and adopt the dataset version and split provided by IBML (Qin et al., 2026).

VGGSound50. VGGSound (Chen et al., 2020) is a large-scale audio-visual dataset covering more than 300 classes of audio events in real-world scenes. We use VGGSound50, the 50-class subset provided by IBML (Qin et al., 2026), which contains 25,954 training clips and 2,499 test clips. All compared methods are reproduced on this 50-class subset.

For VGGSound50 and AVSBench, we use 3 s audio inputs sampled at 22,050 Hz, repeating shorter clips and applying zero padding when needed. For the visual modality, temporally ordered frames are divided into three intervals.

UPMC Food-101. UPMC Food-101 (Wang et al., 2015) is an image–text dataset for multimodal food classification. It contains approximately 100,000 samples across 101 food classes. Images are collected through Google Image Search, and the text comes from the web pages containing these images. We use the same data split as the baselines and follow the text processing settings of IBML (Qin et al., 2026): text is lowercased and tokenized, represented with fixed 300-dimensional GloVe 6B embeddings, and padded to the longest sequence in each mini-batch.

CMU-MOSI. CMU-MOSI (Zadeh et al., 2016) is a multimodal sentiment analysis dataset containing 2,199 opinion segments from 93 YouTube videos by 89 speakers. Each segment includes audio, visual, and textual information and is annotated with a sentiment intensity score from -3 (strongly negative) to $+3$ (strongly positive). We adapt the trimodal experimental pipeline of InfoReg (Huang et al., 2025), using pre-extracted features and the train, valid, and test splits provided in the feature file. We formulate the task as three-class classification: for a continuous sentiment label y, $y \leq -0.5$ is negative, $-0.5 < y < 0.5$ is neutral, and $y \geq 0.5$ is positive. Samples with a zero label are retained in the neutral class.

B.2 Two-stage Training

Algorithm 1 summarizes the audio–visual training procedure. Each run uses a seed-specific split of the original training set, and the calibration subset is included again when the model is trained from scratch in the second stage.

Algorithm 1 Two-stage RCC training for audio–visual classification

Require: Training set $\mathcal{D}$ and RCC parameters in Section 3.2
Randomly split $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$ (90%) and $\mathcal{D}_{\text{held}}$ (10%).
Stage 1: calibration
Initialize model parameters θ and an optimizer; set $\mathbf{p} = (p_{a\to v}, p_{v\to a}) = (0, 0)$.
for $e = 0, \ldots, 59$ **do**
  if $e \in \{30, 40, 50\}$ **then**
    Evaluate the current model on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{held}}$ without gradients.
    Estimate q_{train} and q_{held} using Eq. (3).
    Update $\mathbf{p}$ using Eqs. (4)–(5).
  end if
  Train θ for one epoch on $\mathcal{D}_{\text{train}}$ using Eq. (14) and the loss schedule below.
end for
Freeze the most recently computed $\mathbf{p}$.
Stage 2: training on the full training set
Reinitialize the same model architecture and optimizer; retain only $\mathbf{p}$ from Stage 1.
for $e = 0, \ldots, 99$ **do**
  Train for one epoch on $\mathcal{D}$ using Eq. (14) and the loss schedule below, with fixed $\mathbf{p}$.
end for

In each stage, the first 30 epochs use $\mathcal{L}_{\text{fus}} + \beta\mathcal{L}_{\text{uni}}$; RCC and AP losses are enabled thereafter. AP is omitted during calibration on CREMA-D and KS, and is enabled in the second stage on all datasets. Reliability weights are recomputed for each training mini-batch. Configuration statistics use evaluation mode while retaining the training data transformations. Samples without a different-class donor in the mini-batch are excluded from the corresponding replacement loss.

Table 7: **RCC hyperparameter settings.**

Dataset	α	$p_{\max}$	β	λ_{rcc}	λ_{ap}
CREMA-D	0.35	0.20	4.5	0.05	0.0125
KS	0.15	0.15	4.5	0.02	0.002
VGGSound50	0.05	0.10	3.5	0.005	0.00025
AVSBench	0.15	0.15	4.5	0.005	0.0005
Food101	0.15	0.15	1.0	0.01	0.001
CMU-MOSI	0.15	0.15	—	0.01	0.10

B.3 Hyperparameter Settings

Table 7 reports the dataset-specific RCC hyperparameters. The table reports the final coefficients multiplying $\mathcal{L}_{\text{ap}}$ in the training objective. We tune hyperparameters through parameter sweeps, varying selected parameters while keeping the remaining settings fixed within each group of experiments.

B.4 Experimental on MOSI Dataset

For textual, audio, and visual modalities, the reliability configuration is $\mathbf{r} = (r_t, r_a, r_v) \in \{0, 1\}^3$, giving eight possible combinations of unimodal correctness. Unimodal predictions are obtained from the shared fusion classifier by masking the other two modalities. We consider six ordered replacement directions. For each $s \to m$, we aggregate the configurations satisfying $r_s = 0, r_m = 1$, allowing either correctness state of the third modality. Writing their total frequency as $q_{\mathcal{S}}^{s\to m}$, we compute:

$$D_{s\to m} = \frac{[q_{\text{held}}^{s\to m} - q_{\text{train}}^{s\to m}]_+}{\max(\Pr_{\text{train}}(r_m = 1), \epsilon)}, \qquad p_{s\to m} = \min(\alpha D_{s\to m}, p_{\max}). \tag{44}$$

Each intervention replaces the source feature with a differently labeled donor while retaining the other two modalities and the original label. The target modality provides the soft reliability weight. Per-direction losses are normalized and summed as in the bimodal formulation, with gradients stopped through donor features and weights. Unimodal anchoring, reverse-direction adoption preservation, and the two-stage training procedure follow the main method.

Table 8: **Test accuracies (%) of the compared methods on five bimodal benchmark datasets.** Results are reported as mean ± sample standard deviation over three random runs. The best and second-best results in each column are boldfaced and underlined, respectively.

Method	CREMAD			KS			AVSBench		
	Audio	Visual	Multi	Audio	Visual	Multi	Audio	Visual	Multi
OGM-GE	53.94 ± 1.35	35.08 ± 1.72	76.65 ± 0.51	49.20 ± 1.48	31.55 ± 0.31	69.39 ± 0.56	74.69 ± 0.57	27.63 ± 1.24	88.16 ± 0.29
MMPareto	66.94 ± 0.89	48.30 ± 6.33	73.61 ± 1.01	56.95 ± 0.81	50.64 ± 2.54	73.34 ± 0.44	81.04 ± 0.66	50.13 ± 5.68	90.93 ± 0.63
AUG	68.41 ± 0.29	67.76 ± 0.91	$\underline{83.31 \pm 0.92}$	55.96 ± 0.87	$\mathbf{62.41 \pm 0.06}$	$\underline{76.28 \pm 0.97}$	81.29 ± 0.48	58.60 ± 0.51	91.25 ± 0.44
DGL	$\underline{69.03 \pm 1.10}$	53.07 ± 2.02	77.08 ± 0.58	50.28 ± 1.44	59.15 ± 0.99	75.32 ± 0.47	81.56 ± 0.24	$\mathbf{60.81 \pm 0.73}$	92.01 ± 0.03
ARL	60.71 ± 1.09	$\mathbf{69.04 \pm 1.49}$	77.15 ± 0.54	49.38 ± 0.35	58.94 ± 0.43	74.99 ± 0.52	81.44 ± 0.28	56.82 ± 0.89	91.21 ± 1.03
InfoReg	68.93 ± 1.76	42.80 ± 2.27	74.13 ± 1.06	56.15 ± 0.57	51.31 ± 0.14	73.73 ± 1.01	$\underline{81.59 \pm 0.67}$	53.00 ± 1.49	91.82 ± 0.30
CPSC	68.69 ± 1.07	60.04 ± 0.55	80.33 ± 0.41	$\mathbf{58.63 \pm 0.60}$	$\underline{61.37 \pm 0.02}$	74.46 ± 0.11	79.94 ± 0.19	59.34 ± 0.49	87.41 ± 0.39
IBML	65.41 ± 1.22	65.99 ± 1.55	79.88 ± 0.47	47.97 ± 0.55	51.95 ± 0.77	69.73 ± 0.64	78.95 ± 0.31	59.27 ± 0.72	$\underline{92.31 \pm 0.17}$
TCMax	67.52 ± 0.89	64.34 ± 1.93	79.57 ± 0.35	54.70 ± 0.60	52.52 ± 0.62	76.15 ± 0.49	69.01 ± 3.83	40.20 ± 1.00	87.05 ± 1.05
Ours	$\mathbf{69.35 \pm 1.75}$	$\underline{68.55 \pm 0.27}$	$\mathbf{84.95 \pm 0.27}$	$\underline{57.85 \pm 0.28}$	59.42 ± 0.83	$\mathbf{76.85 \pm 0.31}$	$\mathbf{82.03 \pm 0.73}$	$\underline{59.88 \pm 0.57}$	$\mathbf{93.28 \pm 0.08}$

Method	VGGSound50			Food101		
	Audio	Visual	Multi	Text	Visual	Multi
OGM-GE	40.49 ± 0.69	21.44 ± 1.16	54.79 ± 0.11	35.36 ± 0.60	81.33 ± 0.14	87.88 ± 0.10
MMPareto	45.04 ± 1.43	35.95 ± 0.46	58.50 ± 0.19	51.16 ± 0.54	84.03 ± 0.22	90.38 ± 0.22
AUG	43.63 ± 0.40	39.81 ± 0.72	58.06 ± 0.08	56.31 ± 0.36	83.71 ± 0.34	$\underline{90.67 \pm 0.19}$
DGL	45.23 ± 0.77	39.89 ± 0.27	59.93 ± 0.70	$\underline{56.63 \pm 0.40}$	83.61 ± 0.35	90.06 ± 0.15
ARL	$\underline{46.21 \pm 0.30}$	38.07 ± 0.73	60.04 ± 0.13	56.50 ± 0.34	84.25 ± 0.62	89.51 ± 0.24
InfoReg	45.01 ± 0.64	36.76 ± 0.20	60.02 ± 0.20	51.84 ± 1.02	$\underline{84.46 \pm 0.13}$	90.54 ± 0.31
CPSC	45.10 ± 0.77	$\mathbf{42.52 \pm 0.76}$	56.67 ± 0.54	51.21 ± 0.41	83.81 ± 0.41	89.57 ± 0.22
IBML	43.82 ± 0.84	39.79 ± 0.40	$\underline{60.55 \pm 0.35}$	55.54 ± 0.26	83.65 ± 0.38	90.44 ± 0.23
TCMax	45.81 ± 0.28	33.61 ± 0.96	$\underline{60.55 \pm 0.67}$	53.16 ± 0.40	83.56 ± 0.19	90.53 ± 0.25
Ours	$\mathbf{46.96 \pm 0.26}$	$\underline{40.36 \pm 0.26}$	$\mathbf{61.97 \pm 0.21}$	$\mathbf{57.09 \pm 0.73}$	$\mathbf{84.65 \pm 0.23}$	$\mathbf{91.25 \pm 0.49}$

C Full Accuracy Results

Table 8 reports the complete results over three runs. RCC achieves the highest mean fusion accuracy on all five bimodal datasets, with standard deviations below 0.5 percentage points. On CREMA-D and AVSBench, it also shows lower variation than the strongest competing method on each dataset. The fusion advantage does not require RCC to lead every unimodal metric: on the four audio–visual datasets, other methods achieve higher visual accuracy, while RCC obtains better fusion results. These findings reinforce the distinction between unimodal capability and fusion performance, and show that RCC's strong fusion results are accompanied by limited variation across the evaluated runs.

Table 9: **Full reliability-configuration distributions underlying the calibration deficits.**

Dataset	**Train (%)**				**Held-out (%)**			
	$A\checkmark V\checkmark$	$A\checkmark V\times$	$A\times V\checkmark$	$A\times V\times$	$A\checkmark V\checkmark$	$A\checkmark V\times$	$A\times V\checkmark$	$A\times V\times$
CREMA-D	52.92	47.06	0.00	0.02	35.37	31.94	15.22	17.46
Kinetics-Sounds	58.64	36.47	2.98	1.91	21.40	19.84	27.18	31.58
AVSBench	71.36	28.64	0.00	0.00	39.42	42.03	7.25	11.30
VGGSound50	67.80	31.39	0.49	0.32	38.38	25.13	16.38	20.12

D Full Reliability Configuration Distributions

Table 9 reports all four reliability configurations on the training and held-out calibration subsets. Across the four datasets, the both-correct configuration becomes less frequent on held-out samples, while the both-incorrect configuration becomes more frequent. Within this broader train–held-out difference, the two configurations where only one modality is correct show distinct patterns. On CREMA-D, KS, and VGGSound50, the visual-correct, audio-incorrect configuration increases, whereas the opposite configuration decreases. AVSBench instead shows positive gaps in both configurations. Thus, the direction and magnitude of the coverage differences vary across datasets. The joint correctness distributions identify which conflict configurations have lower coverage in training predictions and provide the statistics used by RCC to allocate additional replacement supervision.

Table 10: **Conditional utility (%) on pairwise common configurations in CREMA-D.** For each comparison model, we intersect its model-induced $A \times V\checkmark$ and $A\checkmark V \times$ subsets with the corresponding RCC subsets, so both models within each block are evaluated on identical samples. N_a and N_v are the mean sizes of the two intersections, and $U_a^{\cap}$ and $U_v^{\cap}$ are the corresponding fusion accuracies. Results are averaged over paired runs. Bold values indicate the higher utility within each block; comparisons are made only within the same block.

Comparison	Model	N_a	$U_a^{\cap} \uparrow$	N_v	$U_v^{\cap} \uparrow$
Plain vs. RCC	Plain	79.33	**95.90**	75.33	47.77
	RCC		86.54		**78.75**
AUG vs. RCC	AUG	76.50	**92.79**	76.50	65.33
	RCC		84.38		**79.75**
ARL vs. RCC	ARL	69.30	**92.75**	84.00	42.13
	RCC		88.82		**78.74**
DGL vs. RCC	DGL	88.30	**95.49**	72.00	37.47
	RCC		86.12		**79.22**
IBML vs. RCC	IBML	79.00	**93.72**	76.70	61.05
	RCC		86.50		**79.65**

E Pairwise Common-Support Evaluation

To compare fusion behavior on the same samples, we evaluate RCC and each baseline on the intersections of their corresponding reliability configurations. Within each block of Table 10, the two models therefore share both the evaluated samples and their unimodal correctness patterns.

RCC achieves higher $U_v^{\cap}$ in all five comparisons, showing better fusion performance on shared samples where both models predict correctly from vision and incorrectly from audio. The visual-utility improvement therefore persists when configuration membership is controlled within each pair. It is accompanied by lower $U_a^{\cap}$, while the lower of the two reported utilities increases and their difference decreases in every comparison. These results support the intended directional effect of RCC on CREMA-D: the $a \to v$ calibration improves fusion performance in the visual-correct configuration that the baselines use less effectively, with an accompanying trade-off in the audio-correct configuration.

Table 11: **Fusion-head control for reliability-conditioned evidence adoption.**

Method	Multi	$\mathbf{U_a}$	$\mathbf{U_v}$	Asym
Plain (Linear)	80.69	96.10	53.63	42.47
Plain (MLP)	79.08	96.67	48.06	48.61
RCC (Linear)	83.42	89.29	74.30	14.99
RCC (MLP)	84.05	85.62	78.63	6.99

F Fusion Head Control Analysis

To examine whether changing the fusion head is sufficient to improve conditional utility, we compare Plain and RCC with linear and MLP classifiers on concatenated modality features. Plain retains fusion supervision and unimodal anchoring but excludes configuration-guided replacement and adoption preservation. In Table 11, replacing the linear head with an MLP under Plain training lowers fusion accuracy and visual conditional utility while increasing utility asymmetry. Additional nonlinear capacity alone therefore does not improve the observed adoption behavior in this experiment.

RCC improves fusion accuracy and visual conditional utility with both heads. Under the respective model-induced reliability partitions, these improvements accompany lower audio conditional utility and smaller utility differences. The gains with linear fusion show that RCC's benefits are not confined to the MLP architecture. RCC with an MLP achieves the highest fusion accuracy and visual utility among the four settings. These results highlight the importance of training supervision

alongside classifier capacity: in this comparison, the MLP benefits from RCC's configuration-guided supervision, whereas changing the head alone does not resolve the imbalance in conditional utility.

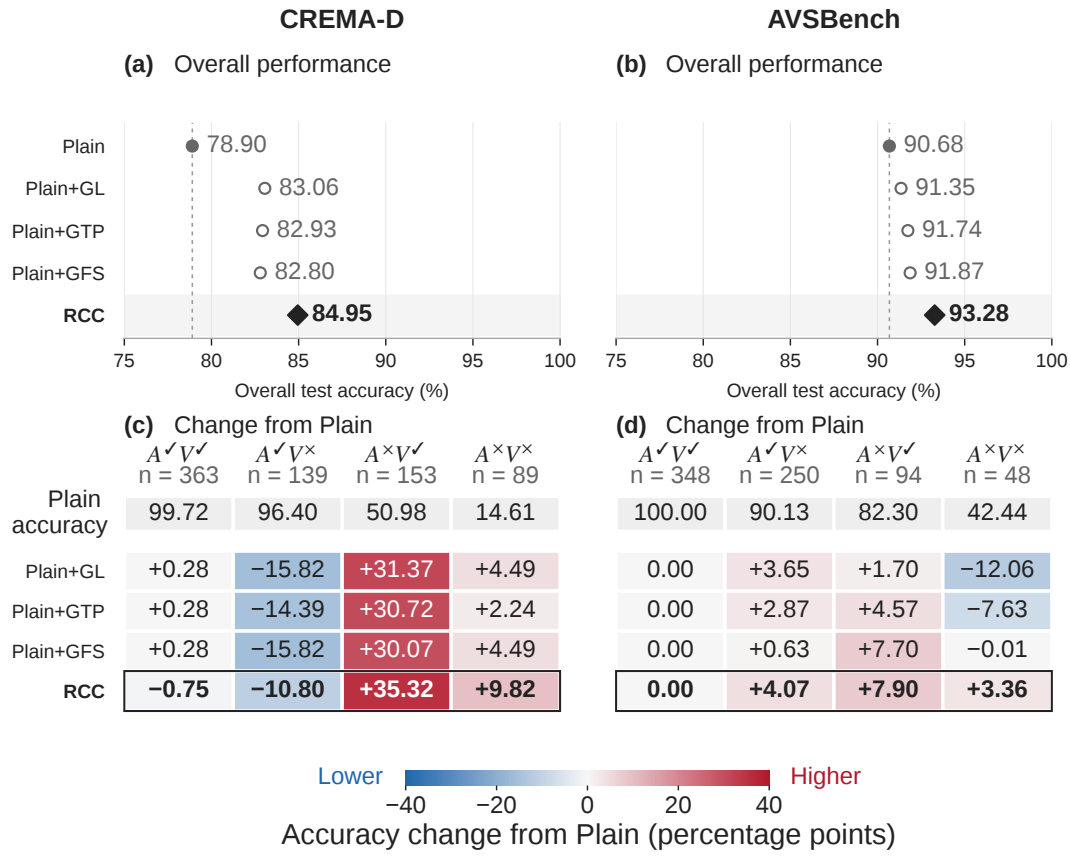


Figure 3: **Comparison with global modality reweighting under fixed reference configurations.** (a,b) Overall test accuracy; dashed lines mark Plain's accuracy. (c,d) Accuracy changes relative to Plain on configurations defined by Plain's masked unimodal predictions. The gray row reports Plain's accuracy, and n denotes the subset size. Colors indicate changes in percentage points.

Table 12: **Fusion learning with shared frozen encoders.** Accuracies are reported in %, and ΔAcc. is the increase over Frozen-Plain in percentage points. U_a^{ref} and U_v^{ref} are accuracies on fixed reference configurations shared across the variants.

Dataset	Method	Multi (%)	ΔAcc.	U_a^{ref}	U_v^{ref}
CREMA-D	Frozen-Plain	80.94	–	80.32	80.46
	Frozen-Plain+Global	81.38	+0.44	79.90	81.45
	Frozen-RCC	82.42	+1.48	83.65	82.12
AVSBench	Frozen-Plain	84.98	–	85.04	79.47
	Frozen-Plain+Global	85.67	+0.69	85.73	82.31
	Frozen-RCC	87.03	+2.05	89.22	85.34

G Controlled Analysis of Fusion Learning

We examine whether RCC's gains can be reproduced by global modality reweighting and whether they persist when encoder representations are held fixed.

G.1 Comparison with Global Modality Reweighting

We compare RCC with three post-hoc variants of Plain: global logit weighting (GL), temperature-scaled probability fusion (GTP), and global feature scaling (GFS) before the original fusion head. All reweighting parameters are shared across samples. The four reference configurations are defined by Plain's masked unimodal predictions and fixed across methods. Global parameters are selected on the official validation set for AVSBench; CREMA-D uses test-selected settings for diagnostic comparison.

Fig. 3 shows that all three global adjustments improve Plain's overall accuracy, while RCC achieves the highest accuracy on both datasets. The configuration-level changes reveal different trade-offs. On CREMA-D, RCC achieves a larger gain on the visual-correct, audio-incorrect subset and a smaller loss on the opposite subset than all three reweighting controls, alongside a small decrease on the both-correct subset. On AVSBench, RCC improves both one-correct subsets and the both-incorrect subset while maintaining accuracy on the both-correct subset. Thus, the global adjustments evaluated here do not fully reproduce RCC's combination of overall accuracy and performance on the one-correct configurations. This is consistent with our motivation to improve the use of correct unimodal evidence under conflicting predictions, rather than relying on a global shift in modality influence alone.

G.2 Fusion Learning with Frozen Encoders

To examine fusion learning with unchanged representations, we freeze a shared pair of encoders and train only the fusion head. Frozen-Plain retains fusion supervision and unimodal anchoring, while Frozen-RCC additionally uses replacement supervision and adoption preservation. Frozen-Plain+Global applies global reweighting to the trained Frozen-Plain model. Evaluation uses a fixed reference partition shared across variants, with U_a^{ref} and U_v^{ref} denoting accuracies on the audio-correct, visual-incorrect and audio-incorrect, visual-correct subsets, respectively.

As shown in Table 12, Frozen-RCC achieves higher overall accuracy and both reference-conditioned utilities than Frozen-Plain on both datasets. Global reweighting also improves overall accuracy, but Frozen-RCC retains its advantage over this control in all three metrics. Because the encoder representations are shared and unchanged, these results indicate that additional fusion supervision can improve decisions made from the same modality features, beyond the global reweighting control evaluated here. This supports our motivation that learning predictive representations and learning to use their evidence effectively are related but distinct objectives.

Table 13: **Counterfactual replacement with soft reliability weighting on CREMA-D.** Statistics are computed on the test set using fixed RCC checkpoints with $\tau = 1$. Target correct and Source wrong are evaluated against the original sample's label; Disagreement measures whether the preserved and replaced modalities predict different classes.

Direction	η	Target correct	Source wrong	Disagreement
$a \to v$	-2	54.87	93.43	87.40
	-1	63.81	93.40	88.62
	0	74.04	93.39	89.99
	1	83.26	93.42	91.24
	2	89.91	93.47	92.16
$v \to a$	-2	86.28	86.27	85.86
	-1	90.60	86.25	85.99
	0	94.51	86.23	86.10
	1	97.25	86.22	86.16
	2	98.76	86.21	86.18

H Counterfactual Replacement Sanity Check

To examine counterfactual replacement and reliability weighting, we construct replacement samples on the CREMA-D test set using fixed RCC checkpoints. We fix $\tau = 1$ and vary η only for diagnosis,

without retraining. Table 13 shows that increasing η places more weight on correct target predictions in both directions, while the weighted source error rate remains approximately unchanged and prediction disagreement stays high. The increase in weighted target correctness is particularly pronounced when vision is preserved. With fixed checkpoints, this change reflects a redistribution of weight toward more reliable target evidence. These observations support the complementary roles of the two components: cross-class replacement frequently supplies a source prediction that is incorrect for the recipient label, while reliability weighting emphasizes correct predictions from the preserved target. The resulting construction provides fusion supervision relevant to the C2U gap, where correct evidence from one modality must be used despite conflicting predictions from another.

Table 14: **Efficiency comparison on CREMA-D.**

Method	Params(M)	Training schedule	Train(sec/epoch)	Infer(sec/total)	Multi
Plain(MLP)	23	100	26.4	7.13	79.08
IBML	22	100	35.5	7.14	79.88
TCMax	22	200	21.6	7.19	79.57
Ours	23	60+100	27.1	7.15	84.95

I Efficiency Analysis

Table 14 compares model size, training time, and inference time on CREMA-D. RCC has the same reported parameter count as Plain and increases the reported per-epoch training time only slightly, from 26.4 to 27.1 seconds. Its additional 60-epoch calibration stage increases the total training cost and is used to determine the replacement probabilities for subsequent training.

At inference, RCC uses the standard fusion prediction without replacement or reliability weighting. Its full-test-set inference time is nearly identical to that of Plain and remains comparable to the other methods. RCC also achieves the highest fusion accuracy in this comparison. These results show that the proposed training procedure improves fusion performance while retaining the model size and inference cost of the underlying fusion network.

J Additional Replacement-Strategy Comparisons

To complement the controlled comparisons in Section 4.6, we evaluate replacement variants on CREMA-D, KS, and AVSBench under the original training protocol. Symmetric replacement (SR) assigns equal probabilities to both directions while preserving their sum. Reverse-direction replacement (RD) swaps the calibrated probabilities. Random replacement (RR) removes class constraints and reliability weights from feature replacement. For each variant, AP probabilities are adjusted together with the corresponding replacement probabilities. SR preserves the total sampling probability; the separately normalized losses need not have equal auxiliary-loss strength(Appendix A.4).

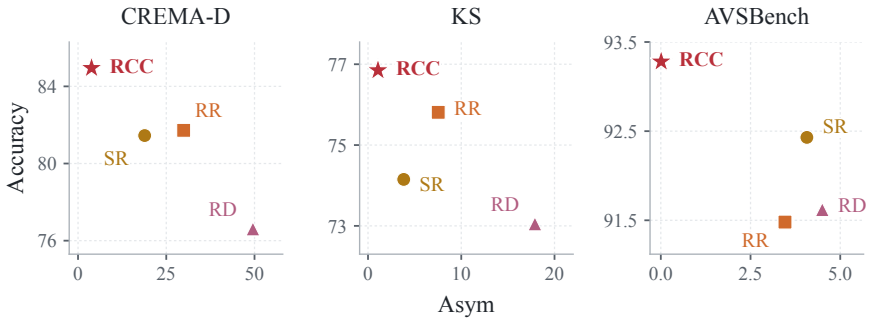


Figure 4: **Additional replacement-strategy comparisons.** Fusion accuracy versus utility asymmetry under the original training protocol. Reliability configurations are defined separately by each model's own unimodal predictions.

Fig. 4 shows that RCC achieves the highest fusion accuracy and the lowest reported utility asymmetry among these variants on all three datasets. RD produces the largest accuracy reductions on CREMA-D and KS, reinforcing the relevance of directional allocation examined by the main-text controls. SR and RR also yield lower fusion accuracy than RCC. The RR comparison additionally supports retaining cross-class donor selection and reliability weighting together when constructing replacement supervision. These results extend the strategy comparison to KS and provide complementary evidence for RCC's combined allocation and sample-construction design.