# If Inputs Conflict, When Does Learning Help?

**Anonymous authors**
Paper under double-blind review

## Abstract

A clear image and a clear audio clip can tell different stories. If one of them belongs to another example, confidence alone cannot tell a multimodal model which story to follow. Learning a correction is an appealing response, but when does it help? We answer this question with a controlled replacement protocol and one default-preserving method, Conflict-Calibrated Evidential Fusion (CEF). CEF keeps the average of source logits as a transparent default, proposes alternative predictions, and accepts one only when a cross-fitted selector predicts a rescue rather than a harm. The method is a measuring instrument as much as a competitor: it lets us score, example by example, when overriding a strong default pays. Three findings trace its benefits and limits. When three of five PolyMNIST views remain target-consistent, coordinate-wise median reaches 97.3% on the official test split, outperforming frozen CEF at 93.6%. With two conflicting sources there is no majority to follow: on AV–MNIST, CEF reaches 54.4% donor accuracy on validation against 48.1% for the best simple rule, and on a reserved test split it raises the default from 48.1% to 55.8% while clean accuracy stays at 96.8%. Finally, a correct answer can be available yet out of reach: with the candidate pool fixed, a selector trained on two replacements reaches only 6.4% on three despite a 99.8% candidate oracle, and retraining on three raises it to 84.3% at a cost of 1.3 points in clean accuracy. The value of learned correction therefore depends on the remaining redundancy, the availability of a correct candidate, and whether the selector generalizes to the corruption regime.

## 1 Introduction

Multimodal models are built on an assumption that is rarely stated and almost never checked at inference time: that every available view describes the same example (Baltrušaitis et al., 2019). We study what happens when it fails silently. In *confident source conflict*, a subset of the input sources is replaced by observations drawn from a different example, while the label and the remaining sources are left intact. The replaced source is not missing, not occluded, and not visibly noisy. It is a clean, high-quality observation of the wrong thing, and it argues for its own class with the same confidence as the sources that are still correct.

This is why the usual defenses do not engage. Confidence carries no information about which source is wrong, because modern networks are confident on inputs they have no business being confident about (Guo et al., 2017). Reweighting schemes that identify an unreliable modality and suppress it (Han et al., 2021; Zhang et al., 2023; Cao et al., 2024) are keyed to the assumption that the unreliable source is the weak one; here the harmful source is the strong one. The consequence is concrete: a single replaced view can move the fused prediction off the correct class while every other view still points at it.

Handling this failure involves two decisions that are usually entangled: constructing a prediction that could recover the target, and judging whether that prediction is better than the current output on this example. Conflict-Calibrated Evidential Fusion (CEF) separates them. Average fusion is the designated default; a donor-trained module constructs correction candidates; and a cross-fitted selector, trained on the comparative event *candi-*



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.

*date correct while default wrong*, admits a candidate only when its predicted gain clears a preservation-constrained threshold.

We use CEF as an instrument rather than as a product. The question we ask is not whether a learned corrector can be made to win, but *when* learning is the right tool at all. Three findings answer it.

**When a majority of sources survives, robust aggregation wins and learning does not help.** On PolyMNIST with five views and two replaced, coordinate-wise median reaches 97.3% on the official test split, against 93.6% for the learned selector, 91.6% for plurality vote, and 76.4% for average fusion (Section 6.1). Robust aggregation is not a strawman in this regime; it is the strongest method in the room, and a learned alternative that cannot beat it should say so rather than change the comparison.

**When no majority exists, the fixed rules lose their footing and learning helps.** On AV–MNIST, where replacing one of two modalities leaves no majority to appeal to, CEF reaches 54.4% on the validation split against 48.1% for the strongest fixed rule we tested; on the reserved test split it raises accuracy from 48.1% to 55.8% (+7.7 points, 95% CI $[5.1, 10.2]$) while clean-input accuracy stays at 96.8% (Figure 2b). These two comparisons are not interchangeable: the first is against simple rules on validation; the second is against the designated default on held-out test. A locked PolyMNIST protocol reproduces the same pattern, 84.5% $\rightarrow$ 90.6%, and applying the same candidate to every example instead yields 82.0%, below the default it replaced (Table 3; Figure 2c).

**The selector is governed by how many sources are corrupted, not by whether a good candidate exists.** Retraining the selector at the harder corruption level on PolyMNIST recovers 84.3%, where a candidate oracle would reach 99.8%: usable candidates are present, and the selector cannot identify them (Figure 2a). Clean accuracy falls by only 1.3 points to 98.7%, so the shortfall is a failure of arbitration rather than indiscriminate routing.

Our contributions are: (i) a protocol for confident source conflict with a same-label benign control, which isolates conflict from mere perturbation; (ii) a default-preserving arbitration framework that separates candidate construction from candidate admission, with a preservation-constrained threshold and cross-fitted training; and (iii) an empirical boundary, replicated on a held-out test split, between the regime where a fixed robust rule suffices and the regime where per-example arbitration is required.

## 2 Related Work

**Robustness to corrupted inputs.** Corruption benchmarks treat degradation as a visible or statistical property of the input—noise, blur, compression, distribution shift (Hendrycks & Dietterich, 2019; Liang et al., 2021). A donor source is drawn from the same clean distribution as its neighbors, so quality-tuned detectors see nothing, and the corruption is legible only in source disagreement. Missing-modality methods train for absent sources (Havaei et al., 2016; Neverova et al., 2016; Ma et al., 2022; Lee et al., 2023); a donor source is present and complete, so it never triggers their fallback.

**Robust multimodal fusion.** A large family estimates per-source reliability and down-weights the unreliable ones—evidential uncertainty (Han et al., 2021; 2023; Zhang et al., 2023), belief discounting (Bezirganyan et al., 2025), adversarial hardening (Yang et al., 2021), predictive dynamic weighting (Cao et al., 2024). A donor source carries full signal about the wrong label, making it harmful in proportion to its quality, and reliability keyed to uncertainty or input degradation has no handle on it. What separates these methods from ours is not the aggregation family but the decision: each commits to one fusion rule for every example; we keep a designated default and decide per example whether to leave it.

Most of this literature assumes joint encoder training, as do conflict-aware designs such as Deregnaucourt et al. (2025) and Bekhouche et al. (2026). CEF reads per-source logits and leaves encoders untouched, suiting frozen or third-party sources but requiring re-

implementation of those methods under an interface they were not designed for; we report them as context rather than controlled baselines.

**Selective prediction and learning to defer.** Abstention frameworks learn when a model should decline (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017); learning-to-defer routes an example to a second decision-maker under a jointly trained policy (Madras et al., 2018; Mozannar & Sontag, 2020), with failure-prediction scores (Hendrycks & Gimpel, 2017; Corbière et al., 2019) as the usual signal. CEF is the two-option case with one reorganizing difference: the alternative is a fallible prediction rather than a reliable expert, so a confidently wrong default is worth overriding only when the candidate happens to be right, and a high failure score on the default is worthless alone. This is why our selector is trained on the comparative event rather than on default correctness.

# 3 Method

## 3.1 Problem setup and intervention protocol

Consider a multimodal example $x = (x^{(1)}, \ldots, x^{(M)})$ with target class $y \in \{1, \ldots, K\}$. Each modality encoder $h_m$ and classifier $q_m$ produce a representation $z_m = h_m(x^{(m)})$ and a class distribution $q_m(y \mid x^{(m)})$, and a default fusion rule combines the available source distributions into $p_0(y \mid x)$.

**Donor replacement.** To create controlled conflict we select a replacement subset $S \subseteq \{1, \ldots, M\}$ with $|S| = k$ and substitute $x^{(m)} \leftarrow \tilde{x}^{(m)}$ for every $m \in S$, leaving the remaining sources and the target label unchanged. A donor label $\tilde{y} \neq y$ creates source conflict, while $\tilde{y} = y$ serves as a benign replacement control separating conflict from mere substitution. At inference the model receives no donor label, donor identity, or condition indicator: the decision must come from the observations alone (Figure 1a).

**Objective.** We seek a fused distribution $p(y \mid x)$ that improves prediction under different-label replacement without degrading clean or same-label performance, which motivates retaining a strong default and evaluating each correction on held-out data before deployment.

# 4 Conflict-Calibrated Evidential Fusion

Here, "evidential" refers to the comparison of source-level predictive evidence and candidate outcomes; it does not imply that the method is restricted to, or equivalent to, Dempster–Shafer belief calculus (Dempster, 1967; Shafer, 1976).

## 4.1 Overview

The central question is when a learned rule should override a simple aggregation, which requires answering example by example while separating three usually entangled decisions: constructing a prediction that might recover the target, judging whether it is better than the current output, and doing so without freely overwriting correct defaults.

CEF addresses this through three components (Figure 1). **Average fusion** (Figure 1b) supplies the default—a transparent baseline representing the consensus of available sources, forcing any override to earn itself against a strong reference. **Candidate construction** (Figure 1b) generates alternatives when the default may be wrong: in the primary PolyMNIST instantiation, predictions from the full source set and every non-empty retained-source subset, yielding leave-one- through leave-four-out candidates ($J = 31$); in the bounded-residual variant, a donor-trained residual branch predicting a tanh-limited correction to the default logits. **A cross-fitted rescue selector** (Figure 1c) estimates whether a candidate corrects a default error on the current example, trained on the comparative event *candidate correct while default wrong*, with greater penalties on harmful overrides; its threshold is selected on held-out predictions (out-of-fold for CEF-Single, a disjoint calibration split for CEF-Rel) under an explicit preservation constraint limiting accuracy loss on clean and other registered conditions.

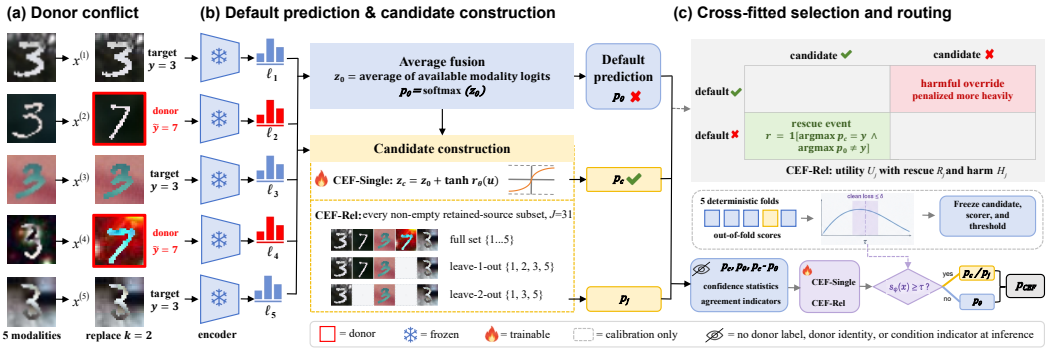



Figure 1: Overview of CEF. (a) Donor replacement: $k = 2$ of the $M = 5$ PolyMNIST views (red boxes) come from a different-label donor, and donor information is hidden at inference. (b) Average fusion gives the default $p_0$; CEF-Single and CEF-Rel construct the candidates $p_c$ and $p_j$. (c) A cross-fitted scorer admits a candidate only if its score clears the calibrated threshold; otherwise $p_0$ is kept.

The framework is a measuring instrument as much as a competitor. Varying which candidates are available, which corruption counts the selector was trained on, and how many target-consistent sources remain measures when learned selection helps and when robust aggregation suffices. Source encoders are frozen throughout, so differences reflect the correction and selection logic rather than representation learning.

### 4.2 Default fusion and bounded single-candidate correction

**Default prediction.** Let $\ell_m \in \mathbb{R}^K$ and $f_m \in \mathbb{R}^d$ denote the logit vector and feature of modality $m$, and let $a_m \in \{0, 1\}$ indicate availability. The default prediction averages the logits of all available sources before applying softmax:

$$z_0 = \frac{\sum_{m=1}^{M} a_m \ell_m}{\sum_{m=1}^{M} a_m}, \qquad p_0 = \mathrm{softmax}(z_0), \tag{1}$$

where $z_0 \in \mathbb{R}^K$ are the averaged logits and $p_0 \in \Delta^{K-1}$ is the resulting class distribution. This serves as the strong default against which corrections compete.

**Bounded residual correction.** Rather than rebuilding the distribution from scratch, the correction branch predicts a bounded adjustment to the default logits. Let $u$ denote the concatenated modality features, logits, and availability indicators. The corrected logits are

$$z_c = z_0 + \tanh r_\theta(u), \qquad p_c = \mathrm{softmax}(z_c), \tag{2}$$

where $r_\theta : \mathbb{R}^D \to \mathbb{R}^K$ is a multilayer perceptron and tanh bounds each correction to $[-1, 1]$. This constraint prevents the branch from freely overwriting the default and forces it to justify any deviation from the consensus of available sources.

The correction is trained to recover donor-corrupted examples while preserving clean predictions:

$$\mathcal{L}_{\mathrm{corr}} = \mathrm{CE}(z_c^{\mathrm{donor}}, y) + \lambda\, \mathrm{KL}(p_0^{\mathrm{clean}} \| p_c^{\mathrm{clean}}), \tag{3}$$

where the superscripts indicate the training condition. The cross-entropy term drives recovery on corrupted inputs; the KL term penalizes deviations from the default on clean examples. The weight $\lambda$ controls the preservation-recovery trade-off. Network architectures and training hyperparameters are detailed in Section 5.1.

### 4.3 Cross-fitted selection and routing

**The rescue event.** In CEF-Single, the correction is not applied unconditionally. We implement the selector as a binary scorer $s_\phi(x) \in [0, 1]$ trained to predict the *rescue event*, defined as the indicator that the candidate corrects a mistake the default makes:

$$r = \mathbf{1}[\arg\max p_c = y \ \wedge \ \arg\max p_0 \neq y]. \tag{4}$$

A rescue occurs when $p_c$ predicts correctly and $p_0$ does not. The scorer estimates $\mathbb{P}(r = 1 \mid x)$ from out-of-fold predictions, trained to distinguish rescue opportunities from harmful overrides—cases where the correction would replace a correct default prediction with a wrong one.

**Scorer design.** The scorer receives features that capture the relationship between the default and candidate predictions: the distributions $p_0$ and $p_c$, their difference, confidence statistics, and agreement indicators. It is trained with asymmetric loss weights that penalize harmful corrections more heavily than missed rescues, reflecting the priority of preserving correct defaults. Cross-fitted training on five deterministic folds generates out-of-fold scores for threshold selection; the final scorer is then refitted on all development predictions (Appendix C and Algorithm C.1 of the supplementary material).

**Routing rule.** Let $\tau \in [0, 1]$ be the selected threshold. The routed prediction is

$$p_{\text{CEF}}(x) = \begin{cases} p_c(x), & s_\phi(x) \geq \tau, \\ p_0(x), & \text{otherwise}, \end{cases} \tag{5}$$

where $\tau$ is chosen to maximize donor accuracy on held-out calibration data subject to a preservation constraint: admissible thresholds may reduce accuracy by at most $\delta$ percentage points on registered clean and low-severity conditions. The threshold trades donor recovery against clean preservation. For AV–MNIST, we additionally constrain clean negative log-likelihood to not increase. The model and threshold are frozen before confirmation inference.

### 4.4 Relational candidate scoring

**Motivation.** When a subset of sources is replaced by donor observations, removing those sources may be more effective than correcting them through a residual. CEF-Rel tests this by constructing predictions from every non-empty source subset and learning to select among them—$J = 2^M - 1$ candidates for $M$ modalities. For PolyMNIST with five modalities we construct the full-source candidate and all leave-one- through leave-four-source-out combinations ($J = 31$ for $M = 5$), with each candidate $j$ averaging the logits of its retained available sources.

**Candidate utility.** To rank candidates during training we define a supervised utility combining prediction quality with rescue-versus-harm accounting:

$$U_j = \text{CE}(z_0, y) - \text{CE}(z_j, y) + \alpha R_j - \beta H_j - \gamma n_j, \tag{6}$$

where $R_j$ and $H_j$ indicate rescue and harm relative to the full-source default, and $n_j$ is the number of removed sources. Higher utility indicates a more valuable candidate: $\alpha < \beta$ penalizes harmful overrides more heavily than missed rescues, and $\gamma > 0$ discourages aggressive source removal unless it provides clear benefit.

**Relational scorer.** The scorer receives a representation of each candidate capturing which sources are retained, which removed, the candidate's predictions, and summary statistics of the retained and removed sets. It scores each candidate and the full-source default, and routes to the highest-scoring candidate $j^\star$ only if its score margin over the default, $\Delta s(x) = s_{j^\star}(x) - s_0(x)$, reaches a preservation-constrained threshold $\tau \geq 0$; the scores are unnormalized logits, so $\Delta s(x)$ is not bounded by 1. The scorer is trained with a listwise loss ranking candidates by their supervised utility $U_j$, combined with a pairwise ranking loss enforcing correctness ordering. Threshold selection and temperature scaling are performed on a disjoint calibration split under constraints that preserve clean accuracy and limit donor harm. Network architectures, feature construction, loss weights, and training hyperparameters are detailed in Section 5.1.

## 5 Experimental Design

### 5.1 Implementation details

All CEF variants use frozen source encoders ($d = 128$ for PolyMNIST), AdamW optimization (lr=$3 \times 10^{-4}$, wd=$10^{-4}$), five-fold cross-fitting for selector training, and threshold selection under preservation constraint $\delta = 0.25$.

**CEF-Single.** Correction networks: PolyMNIST uses $695 \to 256 \to 10$ MLP ($\lambda = 0.25$, 10 epochs); AV–MNIST uses $D \to 64 \to 10$ MLP ($\lambda = 4$, 200 steps). Selectors: PolyMNIST uses gradient boosting (120 iterations, asymmetric weights 4:12 for rescue:harm); AV–MNIST uses $D_g \to 64 \to 1$ MLP (harm weight 8).

**CEF-Rel.** Constructs 31 source-subset candidates on PolyMNIST. Utility: $U_j = \mathrm{CE}(z_0, y) - \mathrm{CE}(z_j, y) + 2R_j - 4H_j - 0.05n_j$. Relational scorer: $361 \to 128 \to 1$ MLP trained with listwise ranking loss, harm weight 4. Threshold selected under an additional harm $\leq 2\%$ constraint.

### 5.2 Benchmarks and interventions

**PolyMNIST** contains five aligned views of each digit. The primary mechanism evaluation replaces two views with views from a different-label donor under the relational candidate-set instantiation; the locked confirmation evaluates the restricted single-candidate instantiation separately. We use five prespecified algorithmic seeds and an independently generated donor assignment, with the locked confirmation adding random-label and same-label donor controls. Every condition uses the same examples for all methods. Appendix Table A.1 of the supplementary material maps every endpoint to its split, seeds, donor assignment, and CEF instance; Appendix D states which claims each supports. **AV–MNIST** pairs audio and image observations. We develop on 1,200 training and 400 validation examples, then evaluate the frozen seed-specific models and thresholds once on a fixed 400-example reserved test split excluded from model selection and threshold tuning. We include **CIFAR10–SVHN** only as a validation-stage transfer boundary; its prespecified criteria are not met, so the official test set is not accessed.

### 5.3 Baselines

**Simple aggregation rules.** We compare against zero-parameter aggregation baselines requiring no training: coordinate-wise median of source logits, plurality vote over source predictions, and average fusion, evaluated row-by-row with CEF in the frozen PolyMNIST sweep (Figure 2a; Table 1). For AV–MNIST we evaluate four fixed two-source selection rules on the validation split—image-only, audio-only, less-confident, and more-confident selection—without reopening the reserved test. Confidence is each source's maximum class probability, and ties resolve deterministically to the lower modality index.

Table 1: PolyMNIST with two replaced sources ($k = 2$), five random seeds. Panel A is the disjoint validation split used for model development; Panel B is the official test split, assessed post hoc with all models, thresholds, and selection decisions frozen beforehand.

| | **Panel A: Validation (1,501)** | | | **Panel B: Official test (10,000)** | | |
|---|---|---|---|---|---|---|
| Method | Acc. (%)↑ | NLL↓ | Brier↓ | Acc. (%)↑ | NLL↓ | Brier↓ |
| Coordinate-wise median | **97.002** | **0.0976** | **0.0480** | **97.300** | **0.0805** | **0.0416** |
| Plurality vote | 91.406 | 0.5765 | 0.3535 | 91.550 | 0.5693 | 0.3479 |
| Average fusion | 75.909 | 0.7030 | 0.3645 | 76.350 | 0.6667 | 0.3448 |
| Frozen CEF-Rel ($k = 2$) | 92.791 | 0.3506 | 0.1377 | 93.576 | 0.3466 | 0.1229 |

**Trainable fusion methods.** We implemented trainable fusion methods for PolyMNIST: product of experts (PoE), Dempster–Shafer fusion with conflict redistribution, discounted belief fusion (DBF) (Bezirganyan et al., 2025), learned attention fusion, and modality-concatenation with an MLP classifier. We retain these as implementation records rather than main-table competitors, since we cannot verify identical donor-corruption training conditions, random seeds, and hyperparameter-selection protocols across them; a controlled comparison would require retraining each under CEF's frozen-encoder, cross-fitted, preservation-constrained protocol, beyond the scope of this boundary study.

**Conflict-aware fusion without learned selection.** To isolate what learned candidate selection contributes, we include a uniform-reliability baseline: conflict-aware discounting with no rescue selector. It uses the evidential fusion operator with $q_m = 1$ (equal prior reliability) for every available source, discounts each source's Dirichlet evidence vector $e_m$

by its attributed pairwise conflict $c_m$, and sums the discounted evidence:

$$d_m = \exp(-\kappa c_m), \qquad e_{\text{unif}} = \sum_{m=1}^{M} a_m d_m e_m, \qquad p_{\text{unif}} = \frac{e_{\text{unif}} + \mathbf{1}}{\mathbf{1}^\top (e_{\text{unif}} + \mathbf{1})}, \tag{7}$$

where $\kappa$ is a fixed discounting rate. With no residual correction, candidate pool, or learned selector, it upper-bounds what conflict-aware weighting alone can achieve.

### 5.4 Metrics and inference

**Evaluation metrics.** We report three primary metrics: **accuracy** for target-class recovery; **negative log-likelihood (NLL)** and **Brier score** (Brier, 1950) for probabilistic calibration. For methods with learned selection we add **route rate** (fraction sent to the candidate), **rescue rate** (fraction corrected from wrong to right), and **harm rate** (fraction changed from right to wrong), with net accuracy gain being rescue minus harm. To separate candidate quality from selector quality we report a **candidate oracle**: the fraction of examples for which at least one candidate in the frozen pool predicts correctly. High oracle with low accuracy means useful candidates exist but the selector cannot identify them; low oracle means the pool is insufficient.

**Statistical inference.** The independent unit of replication is the algorithmic seed, not the individual prediction. We train and evaluate five models with different random seeds, each yielding one summary statistic over all examples, and report arithmetic means over the five seed-level statistics, which captures model variability without pseudoreplication. For the AV–MNIST reserved-test confirmation, the 95% confidence interval is a two-sided paired Student-$t$ interval over the five seed-level differences ($df = 4$), conditional on the fixed donor assignment.

**Sensitivity to donor assignment.** To check that results are not driven by a single donor assignment, we ran a three-draw sensitivity check on the frozen PolyMNIST $k = 2$ evaluation. Across three independently generated donor assignments, frozen CEF-Rel achieved 92.791%, 93.418%, and 93.205% (mean 93.138%, s.d. 0.318 points, range 0.63); the median baseline was similarly stable (s.d. 0.298 points), and the median–CEF gap held at 3.6 points across the sensitivity draws, 4.2 on the first preregistered assignment, and 3.7 on the official test split. The median advantage is therefore systematic rather than assignment-specific. Accuracy intervals are in percentage points unless noted.

## 6 Results

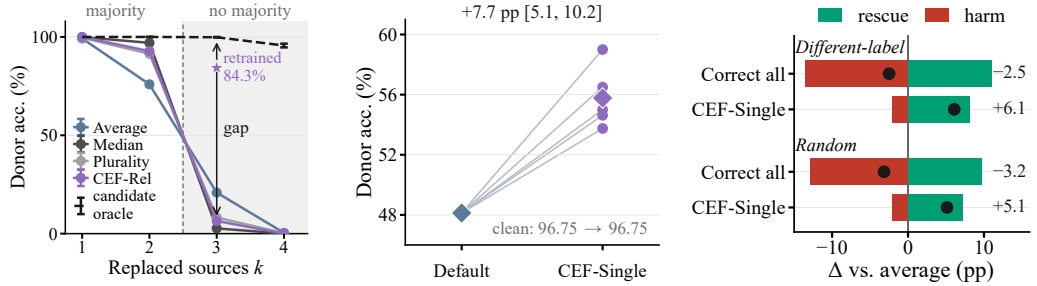
### 6.1 When a majority remains, robust aggregation is the right default

We begin with the result that locates the boundary: when enough clean sources remain, a simple robust aggregator outperforms the tested learned correction. The frozen PolyMNIST $k$-sweep (Protocol P1; Appendix Table A.1) varies the number of replaced sources from one to four while keeping the CEF-Rel selector frozen at its $k = 2$ training point (Figure 2a).

**At $k = 2$, median outperforms frozen CEF.** With two of five views replaced, coordinate-wise median achieves 97.0% mean accuracy on the validation split against 92.8% for frozen CEF-Rel, 91.4% for plurality vote, and 75.9% for average fusion (Table 1, Panel A)—with no training and no parameters. On the official test split the same frozen pipeline yields 97.3% for median and 93.6% for CEF-Rel (Panel B), a median advantage of 3.6–4.2 percentage points across splits and donor assignments; median also attains the lowest NLL and Brier score on both splits. In this redundant-view setting—three clean sources forming a strict majority—learning to correct the default does not surpass robust aggregation.

**At $k = 3$, the majority assumption breaks.** Once three of five views are replaced, the honest views no longer form a majority, and every frozen rule collapses: average fusion, median, plurality, and frozen CEF-Rel reach only 20.9%, 2.7%, 8.2%, and 6.4% (Figure 2a). The $k$-sweep identifies a transition from a redundancy-supported regime, where aggregation suffices, to a no-majority regime where selection becomes the central problem—the two-source setting of Section 6.2, where no strict majority exists.

**Role of learning.** CEF-Rel recovers most of what average fusion loses, lifting accuracy from 75.9% to 92.8%, but does not close the gap to median. It complements robust aggregation rather than replacing it: when the majority assumption holds, median is the right default, and the following sections examine settings where it no longer applies.




(a) PolyMNIST corruption sweep (b) AV–MNIST reserved test (c) Rescue–harm decomposition

Figure 2: Main results (five-seed means). (a) PolyMNIST corruption sweep with correction and selector frozen at $k = 2$ (P1); shading marks the no-majority regime, the dashed line is the candidate oracle, and the star is the selector retrained at $k = 3$ (P3); error bars are s.d. (b) AV–MNIST reserved test (P5); thin lines pair default fusion and CEF-Single per seed (all seeds share the default), diamonds are means. (c) Rescue (green) and harm (red) relative to average fusion on locked PolyMNIST (P2); dots are net change.

### 6.2 With two sources, the gain is not a fixed preference

Table 2: AV–MNIST, one source replaced. Panel A: reserved test, evaluated once; Panel B: validation. Panels are not directly comparable.

| Method | Clean (%) | Donor (%) |
|---|---|---|
| **Panel A: Reserved test** | | |
| Default fusion | 96.750 | 48.125 |
| CEF-Single | 96.750 | **55.775** |
| **Panel B: Validation** | | |
| Image-only | 81.250 | 42.000 |
| Audio-only | 95.000 | 48.125 |
| Less-confident | 79.500 | 42.250 |
| More-confident | 96.750 | 48.125 |
| CEF-Single | n/a | **54.425** |

Table 3: Routing ablation on locked PolyMNIST (P2), official test split, five-seed means (%). Correction applies the residual to every example; CEF-Single routes only above the calibrated threshold.

| Condition | Average | Correction | CEF-Single | Route | Harm |
|---|---|---|---|---|---|
| Clean | 99.698 | 100.000 | 100.000 | 0.302 | 0.000 |
| Different-label | 84.528 | 82.038 | **90.610** | 10.398 | 2.054 |
| Random | 86.182 | 83.020 | **91.320** | 9.374 | 2.012 |
| Same-label | 99.728 | 100.000 | 100.000 | 0.272 | 0.000 |

With only two sources, no strict majority exists to break the tie. On AV–MNIST we evaluate CEF-Single with one modality replaced against simple two-source selection rules. Table 2 reports two complementary evaluations on distinct splits: Panel A confirms the improvement over the default on a reserved test set never used for development; Panel B tests whether that gain reduces to a fixed preference by comparing against simple baselines on the validation split.

**Reserved-test confirmation (Panel A).** The frozen CEF-Single model raises donor accuracy from 48.1% to 55.8% on the 400-example reserved test split, an improvement of +7.7 percentage points (95% CI $[5.1, 10.2]$, paired $t$-test over five seeds), with every seed improving (Figure 2b) and clean accuracy unchanged at 96.8%. The split was held out during all model development and threshold selection, with the model and threshold frozen before this single evaluation.

**Ruling out simple explanations (Panel B).** To test whether the gain comes from always trusting one modality or following a fixed confidence ranking, we compare CEF-Single against four simple two-source rules on the validation split: image-only (42.0%), audio-only (48.1%), less-confident (42.3%), and more-confident (48.1%). The strongest simple rule matches the default but falls short of CEF-Single at 54.4% on the same split, ruling out a fixed modality preference or confidence ordering.

### 6.3 The selector is conditioned on the corruption regime

The frozen $k$-sweep (Figure 2a) showed that when three of five sources are replaced, the $k = 2$-trained selector falls to 6.4% accuracy, although the candidate oracle stays at 100.0% for $k \leq 2$, 99.8% at $k = 3$, and 95.6% at $k = 4$. To determine whether the candidate pool or the selector is responsible, we retrained only the selector on $k = 3$ data (Protocol P3), holding the candidate generator, 31-candidate pool, source predictions, and evaluation examples fixed.

With the pool unchanged, retraining on $k = 3$ raises donor accuracy from 6.4% to 84.3% and lowers donor NLL from 4.99 to 0.86 (star in Figure 2a). The 78-point gap isolates the selector as the failure mode: the pool contains the needed answers, but the $k = 2$-trained selector cannot identify them. Retraining reduces clean accuracy by 1.3 points to 98.7%, so it is evidence about the mechanism rather than a zero-cost fix. The 6.4% failure at $k = 3$, combined with the oracle–actual gap, shows that the selector is conditioned on the corruption-count distribution it saw during training: it has learned patterns specific to the $k = 2$ setting, and those patterns do not apply when the counts reverse. A cross-configuration audit (Appendix Table A.2) separates candidate availability from selector quality in both directions: replacing the relational scorer lowers accuracy despite an unchanged oracle, and attaching it to the single-candidate generator raises accuracy.

### 6.4 Selective routing, rather than unconditional correction, provides the gain

CEF-Single improves over the default on AV–MNIST, and the locked PolyMNIST confirmation lets us isolate the source of that gain: is the correction candidate itself better, or does the value come from knowing when to use it?

Table 3 reports the locked PolyMNIST confirmation (Protocol P2; Appendix Table A.1) on the official 10,000-example test split. Three accuracy columns isolate the routing component: Average is the default fusion; Correction applies the bounded-residual correction to every example without selection; CEF-Single routes to the correction only when the selector score exceeds the calibrated threshold. On different-label donor conflict, unconditional correction achieves 82.0%, below the default at 84.5%, whereas selective routing reaches 90.6% by routing only 10.4% of examples; on clean and same-label inputs it routes fewer than 0.4% of examples and causes no harm. Figure 2c decomposes this effect relative to average fusion. Unconditional correction rescues 11.0% of examples but harms 13.5%, a net change of $-2.5$ points; CEF-Single rescues fewer (8.1%) but cuts harm to 2.1%, a net gain of $+6.1$ points, and random donors show the same pattern ($-3.2$ versus $+5.1$ points). The correction candidate is not uniformly better; the value comes from recognizing when to apply it. Threshold selection behaves the same way for CEF-Rel: on the PolyMNIST development calibration split (P11; Appendix Figure B.1), rescue exceeds harm and clean accuracy is unchanged at every one of the ten prespecified score-margin thresholds, so the registered 0.25-point preservation constraint never binds—on this split, the scorer's margin separates rescues from harms at every threshold tested.

## 7 Conclusion

Robust multimodal conflict resolution shows two regimes in the settings we study: fixed robust aggregation is preferable while an honest majority remains; selective learned correction becomes useful once that structural advantage disappears. CEF implements the latter by preserving a default prediction and routing to a selected alternative, but its selector remains conditioned on the corruption regime used in training. The contribution is a mechanism and a boundary: candidate availability and candidate selection are separate bottlenecks, and improving one does not guarantee transfer of the other. The central limitation is distribution dependence—the main PolyMNIST mechanism results are validation-stage analyses, and selector-only retraining improves the $k = 3$ donor result but costs 1.3 points on clean inputs. Making the selector transfer across corruption mechanisms without sacrificing default preservation is the next challenge.

## AI Use Statement

We used a large language model to assist with language editing, checking bibliographic entries against their sources, drafting prompts for figure layout, and formatting LaTeX tables and figures. The research questions, method, experiments, results, and conclusions are the authors' own. All AI-assisted text and references were verified by the authors, who take full responsibility for the content of this paper.

## References

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

S. E. Bekhouche, H. Telli, A. Benlamoudi, S. E. Herrouz, A. Taleb-Ahmed, and A. Hadid. Conflict-aware multimodal fusion for ambivalence and hesitancy recognition. *arXiv preprint arXiv:2603.15818*, 2026.

G. Bezirganyan, S. Sellami, L. Berti-Équille, and S. Fournier. Multimodal learning with uncertainty quantification based on discounted belief fusion. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3142–3150, 2025.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

B. Cao, Y. Xia, Y. Ding, C. Zhang, and Q. Hu. Predictive dynamic fusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5608–5628, 2024.

C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967. doi: 10.1214/aoms/1177698950.

L. Deregnaucourt, H. Laghmara, A. Lechervy, and S. Ainouz. A conflict-guided evidential multimodal fusion for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1373–1382, 2025.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (2):2551–2566, 2023. doi: 10.1109/TPAMI.2022.3171983.

M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. HeMIS: Hetero-modal image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 469–477, 2016. doi: 10.1007/978-3-319-46723-8_54.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952, 2023.

P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, R. Salakhutdinov, and L.-P. Morency. MultiBench: Multiscale benchmarks for multimodal representation learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

D. Madras, T. Pitassi, and R. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020.

N. Neverova, C. Wolf, G. Taylor, and F. Nebout. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8): 1692–1706, 2016. doi: 10.1109/TPAMI.2015.2461544.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

K. Yang, W.-Y. Lin, M. Barman, F. Condessa, and Z. Kolter. Defending multimodal fusion models against single-source adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3340–3349, 2021. doi: 10.1109/CVPR46437.2021.00335.

Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769, 2023.