PROVABLY ROBUST MULTIMODAL FUSION UNDER CONDITIONAL DEPENDENCE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal robustness is usually studied by corrupting or degrading individual modalities. We study a different failure mode where the quality distribution of each modality remains unchanged while the dependence among their local quality states changes. We call this *conditional dependence shift* (CDS). Under CDS, every modality experiences the same corruption marginal, yet fusion can still fail because local failures co occur differently across modalities, changing the temporal overlap of jointly reliable evidence. We characterize exactly when fusion risk is invariant to such quality couplings, derive an exact decomposition of the resulting risk shift, and introduce a coupling-sensitivity modulus that tightly bounds risk variation under every coupling with the same modality marginals and is computable by a small linear program. These results yield StableFuse, which aggregates evidence and reliability within each modality and directly regularizes a differentiable upper bound on this modulus, so that robustness to coupling shifts is explicitly controlled rather than only encouraged. Using controlled within sample protocols that preserve individual modality quality while changing only the coupling of local failures, StableFuse improves average and worst case CDS performance across four multimodal benchmarks while reducing the certified modulus and preserving clean accuracy.

1 INTRODUCTION

Multimodal learning improves prediction by combining complementary information from images, text, audio, and other sources. Its reliability, however, depends strongly on modality quality. A large body of work therefore studies multimodal prediction when one or more modalities become unreliable. Quality aware fusion adjusts modality contributions according to estimated reliability (Zhang et al., 2023). Uncertainty based fusion suppresses evidence with high aleatoric uncertainty (Gao et al., 2024), while conformal predictive calibration connects modality corruption and optimization imbalance through predictive uncertainty (Jiang et al., 2026). These approaches improve robustness to noisy, incomplete, or low quality modalities (Zhang et al., 2024; Hu et al., 2024; Kim & Kim, 2024; Ke et al., 2025; Zhuang et al., 2026).

Recent work has further shown that multimodal robustness depends on more than individual modality quality. Multimodal models can exploit correlations that do not remain stable across environments. Prior studies mitigate spurious correlations during multimodal fine tuning (Yang et al., 2023; Lu et al., 2025; Li et al., 2025), analyze multimodal contrastive learning under distribution shift (Xue et al., 2024), and improve cross modal alignment under correlation and diversity shifts (Zhu et al., 2023). Test time adaptation has also been studied under modality specific and simultaneous multimodal shifts (Yang et al., 2024; Dong et al., 2025; Chen et al., 2025; Guo & Jin, 2025; Li & Feng, 2026). These studies show that cross modal relationships can themselves become a source of failure.

Yet existing robustness settings typically change the quality or reliability of individual modalities. In real multimodal systems, however, local reliability changes need not occur synchronously. A video stream may temporarily become stale because of buffering while audio remains current, packet loss may affect different streams at different times, and sensor specific latency can alter which modalities are reliable simultaneously. Crucially, such changes need not increase the marginal failure rate, duration, or severity of any individual modality. They can instead change how local failures co occur across modalities. This raises a different question that remains largely unexplored: if each modality

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

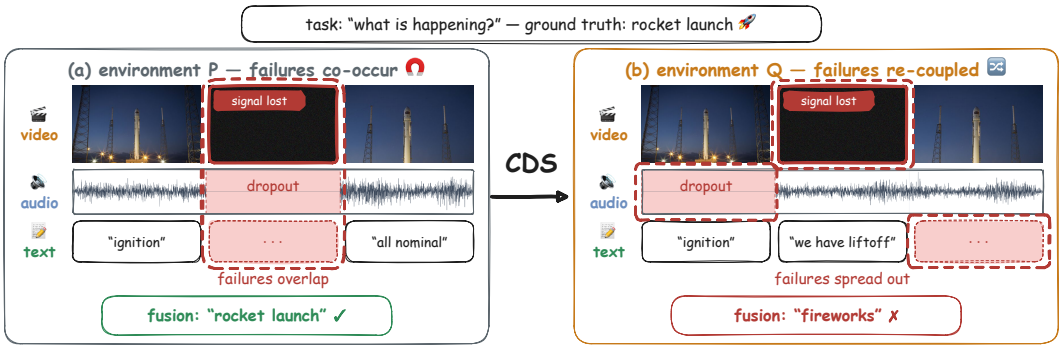

Figure 1: Conditional dependence shift preserves clean sample content and individual modality quality while changing how local failures co occur across modalities. The same amount of corruption can lead to different fusion behavior when the failure coupling changes.

has the same quality distribution across environments, *can fusion still fail because their local failures occur together in a different way?*

As illustrated in Fig. 1, the answer can be yes. When local failures overlap across modalities, corruption is concentrated in fewer temporal regions and the remaining regions retain more jointly reliable evidence. When the same failures occur at different regions, the amount of corruption in every modality remains unchanged, but the temporal overlap of reliable evidence decreases. A fusion model that relies on cross modal interaction can therefore behave differently even though no individual modality becomes less reliable. In short, stable marginal quality does not imply stable fusion.

We formalize this phenomenon as conditional dependence shift (CDS). Unlike conventional corruption shifts, CDS keeps the clean sample content and the quality distribution of each individual modality fixed. What changes is how local quality states are coupled across modalities. This isolates a source of multimodal instability that cannot be attributed to any modality becoming individually less reliable.

We ask what makes a fusion predictor sensitive to such changes, and whether this sensitivity can be measured and controlled. We show that coupling sensitivity is governed by cross modal interaction among quality states in the loss, and derive an exact decomposition of the risk shift. We introduce a coupling-sensitivity modulus, the distance from the loss to the class of additive quality effects, which tightly bounds risk variation under couplings with the same modality marginals and is computable by a linear program. CDS robustness thus becomes an explicit, optimizable quantity.

This characterization yields **StableFuse**. StableFuse aggregates evidence and reliability within each modality, and trains against the theory: it assembles the loss tensor over joint quality states from per-bin features at negligible cost and regularizes its interaction residual, a differentiable upper bound on the modulus. Risk variation under coupling shifts is thereby explicitly controlled rather than only discouraged, while clean and per-modality objectives preserve complementary evidence. We evaluate CDS through controlled within sample perturbations that abstract asynchronous reliability patterns by fixing the amount and severity of local corruption within every modality while varying only the co occurrence of failures. Across four multimodal benchmarks, we report clean, average, and worst case coupling performance, alongside the certified modulus with and without interaction regularization.

Our contributions are summarized as follows.

- We identify conditional dependence shift as a distinct multimodal failure mode where clean content and individual modality quality distributions remain fixed while the dependence among local quality states changes.
- We derive an exact risk decomposition for CDS and introduce a computable coupling-sensitivity modulus that certifies risk variation under every quality coupling with fixed modality marginals.
- We propose StableFuse, which combines modality local evidence aggregation with quality interaction regularization, directly minimizing a differentiable upper bound on the certified modulus while preserving complementary evidence across four multimodal benchmarks.

2 Related Work

Robust multimodal fusion. Robust multimodal learning has been studied under modality imbalance, noisy inputs, missing modalities, and modality reliability Wang et al. (2020); Peng et al. (2022); Fan et al. (2023); Wei & Hu (2024); Zhang et al. (2023); Gao et al. (2024); Jiang et al. (2026); Ma et al. (2022); Woo et al. (2023); Lin & Hu (2023). Related methods improve robustness through adaptive weighting, uncertainty estimation, gradient modulation, calibration, and mixture based fusion Zhang et al. (2023); Gao et al. (2024); Jiang et al. (2026); Wei & Hu (2024); Xin et al. (2025). These works mainly address changes in the quality, availability, or reliability of individual modalities. In contrast, CDS keeps the clean content and individual modality quality distributions fixed while changing how local quality states are coupled across modalities.

Multimodal distribution shift and interaction. Multimodal robustness has been studied under spurious correlations, domain shifts, correlation shifts, and test time changes Yang et al. (2023); Xue et al. (2024); Zhu et al. (2023); Yang et al. (2024); Arjovsky et al. (2019); Sagawa et al. (2019); Krueger et al. (2021); Gulrajani & Lopez-Paz (2020); Cha et al. (2021). Another line of work analyzes cross modal interactions, redundancy, uniqueness, and synergy in multimodal prediction Liang et al. (2023; 2021); Tsang et al. (2017). In particular, information decomposition considers distributions preserving selected marginal relationships Liang et al. (2023). Our setting targets a different source of instability. We preserve every modality's quality marginal, changing only their joint dependence, and characterize when such changes can alter fusion risk through cross modal quality interaction.

Distributionally robust learning. Distributionally robust optimization improves worst case performance over ambiguity sets defined by divergences, moments, or transportation costs Ben-Tal et al. (2013); Namkoong & Duchi (2016); Sinha et al. (2017); Gao & Kleywegt (2023); Mohajerin Esfahani & Kuhn (2018); Blanchet & Murthy (2019); Duchi et al. (2021); Lin et al. (2022). Related work in optimal transport and extremal dependence studies joint distributions with prescribed marginals Courty et al. (2016); Rüschendorf (1982); Puccetti & Wang (2015). Our problem brings this fixed marginal perspective to multimodal quality dependence. Rather than only optimizing worst case risk over the coupling family, StableFuse directly regularizes the interaction term that makes risk coupling-dependent, with the certified modulus bounding the resulting risk variation.

3 From Conditional Dependence Shift to StableFuse

3.1 Stable marginal quality, unstable dependence

We study a multimodal failure mode in which every modality has the same quality distribution across environments, while the way local failures occur together changes. Consider an audio visual sample divided into temporal regions, where a local sensor stall makes a modality repeat its last reliable observation. Audio and video may fail at the same region in one environment but at different regions in another: failure duration and severity are unchanged, yet jointly reliable evidence changes. When failures overlap, corruption concentrates in fewer regions and the rest retain more jointly reliable evidence; when failures disperse, reliable evidence overlaps less across modalities. Fusion can therefore behave differently even when no individual modality becomes less reliable.

We formalize this distinction by separating clean content from local quality. Let $C = (X_1, \dots, X_M)$ denote the clean multimodal content and $S = (S_1, \dots, S_M)$ denote the modality quality states. The observed modality is $\widetilde{X}_m = T_m(X_m, S_m)$, where T_m describes how quality state S_m affects modality m. For fixed clean content c and label y, let $\mu_{m,c,y}$ denote the conditional quality marginal of modality m, and let $\mathcal{C}_{c,y}$ denote the set of joint quality distributions sharing these marginals. Conditional dependence shift (CDS) changes the quality coupling inside this set while preserving the clean content distribution:

$$P_0(C, Y) = Q(C, Y), \qquad \nu^0_{c,y}, \nu^Q_{c,y} \in \mathcal{C}_{c,y}, \qquad \nu^0_{c,y} \neq \nu^Q_{c,y}, \tag{1}$$

where $\nu^0_{c,y}$ and $\nu^Q_{c,y}$ are the conditional joint distributions of S in the two environments. Neither clean content nor individual modality quality changes under this definition, and only the dependence among quality states is allowed to vary. CDS therefore isolates the instability caused by quality coupling rather than by stronger unimodal corruption.

3.2 QUALITY INTERACTION GOVERNS CDS SENSITIVITY

We consider finite quality state spaces throughout. If every individual quality distribution remains fixed, what property of fusion allows the joint coupling alone to change its risk? Fix clean content c and label y, and let $L_{c,y}(s)$ denote the loss at joint quality state $s = (s_1, \ldots, s_M)$. Let $\mathcal{H}_{\text{add}}$ denote the class of additive quality effects,

$$\mathcal{H}_{\text{add}} = \left\{ g(s) = g_0 + \sum_{m=1}^{M} g_m(s_m) \right\}. \tag{2}$$

Using the product of the fixed quality marginals as the reference measure, we decompose the loss as

$$A_{c,y} = \text{Proj}_{\mathcal{H}_{\text{add}}} L_{c,y}, \qquad J_{c,y} = L_{c,y} - A_{c,y}. \tag{3}$$

The additive component $A_{c,y}$ captures individual modality quality effects, while the residual $J_{c,y}$ captures the remaining cross modal quality interaction: a nonzero $J_{c,y}$ means the effect of one modality quality state depends on the simultaneous quality states of other modalities, and cannot be attributed to any single modality. The explicit projection formula is given in the appendix.

When can quality coupling matter? If the loss is fully explained by separate modality quality effects, changing how these states occur together cannot change its expectation.

Theorem 1 (Quality coupling sensitivity)**.** Assume that every quality marginal has full support. For fixed c and y, the expected loss is invariant to every $\nu \in \mathcal{C}_{c,y}$ if and only if $J_{c,y}(s) = 0$ for every valid joint quality state s.

CDS sensitivity therefore does not arise merely because observations are degraded; it appears when the loss responds to how quality states occur together, which is why overlapping and dispersed failures can behave differently under the same amount of corruption.

How does quality coupling change risk? Since $\nu^0_{c,y}$ and $\nu^Q_{c,y}$ share the same modality marginals, all additive quality effects cancel, and the risk shift acts entirely through the interaction.

Theorem 2 (CDS risk decomposition)**.** For every Q satisfying equation 1,

$$R_Q(f) - R_{P_0}(f) = \mathbb{E}_{(C,Y)\sim P_0}\left[\left\langle \nu^Q_{C,Y} - \nu^0_{C,Y}, J_{C,Y} \right\rangle \right]. \tag{4}$$

Risk changes when probability moves between joint quality states with different interaction effects.

A computable sensitivity certificate. The decomposition also yields a quantitative certificate. We measure the distance from the loss to the entire additive class,

$$\Omega_{c,y}(f) = \inf_{A \in \mathcal{H}_{\text{add}}} \|L_{c,y} - A\|_\infty, \tag{5}$$

which we call the *coupling-sensitivity modulus*. Since the projection $A_{c,y} \in \mathcal{H}_{\text{add}}$ is one admissible choice, we always have $\Omega_{c,y}(f) \le \|J_{c,y}\|_\infty$.

Theorem 3 (Coupling-sensitivity certificate)**.** Assume every quality marginal has full support. Then $\Omega_{c,y}(f) = 0$ if and only if $J_{c,y} = 0$. Moreover, for any two couplings $\nu, \nu' \in \mathcal{C}_{c,y}$,

$$|\mathbb{E}_\nu[L_{c,y}] - \mathbb{E}_{\nu'}[L_{c,y}]| \le \|\nu - \nu'\|_1 \, \Omega_{c,y}(f) \le \|\nu - \nu'\|_1 \, \|J_{c,y}\|_\infty. \tag{6}$$

Consequently, if $\|\nu^Q_{c,y} - \nu^0_{c,y}\|_1 \le \rho$ for every c and y, then $|R_Q(f) - R_{P_0}(f)| \le \rho \, \mathbb{E}_{(C,Y)\sim P_0}[\Omega_{C,Y}(f)]$.

On finite state spaces, $\Omega_{c,y}$ is the value of a small linear program (appendix), so the certificate is computed exactly for evaluation, while $\|J_{c,y}\|$ provides a differentiable upper bound for training. Since the loss tensor depends on the label, $\Omega_{c,y}$ is an evaluation-time diagnostic; a label-free variant $\bar{\Omega}_c = \max_y \Omega_{c,y}$ is discussed in the appendix. Stable fusion should therefore preserve useful modality evidence while driving this interaction toward zero.

3.3 STABLEFUSE

StableFuse instantiates this principle at two levels: an architecture that aggregates modality-local evidence, and an objective that directly regularizes the interaction identified by Theorem 3.

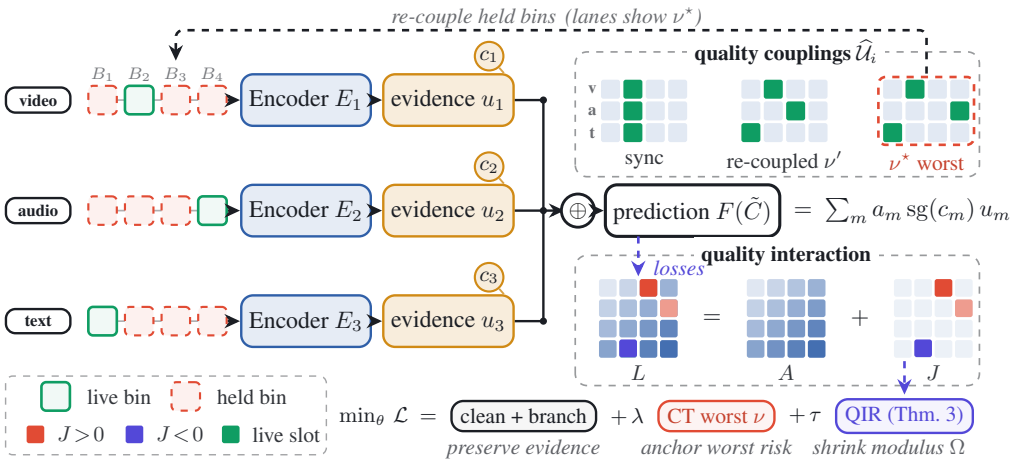

Figure 2: Overview of StableFuse. Modality-local evidence is fused additively. Admissible couplings $\widehat{\mathcal{U}}_i$ change only the cross-modal arrangement of held bins with unimodal marginals fixed. The worst coupling in Eq. (9) is $\nu^\star$, and QIR penalizes the interaction J behind the CDS risk shift.

Modality-local evidence. Predictive evidence is extracted independently from each modality before cross modal aggregation,

$$u_m = E_m(\widetilde{X}_m), \qquad F(\widetilde{C}) = \sum_{m=1}^{M} a_m c_m u_m, \tag{7}$$

where E_m never inspects another modality's quality state or representation, a_m is a positive normalized scale, and c_m is a confidence computed from u_m alone. This removes explicit predictors of joint failure configurations while retaining evidence from all modalities. Because the classification loss is nonlinear, additivity alone does not make $J_{c,y}$ vanish, but it provides an architecture on which the interaction can be driven small.

Quality interaction regularization. We therefore regularize the interaction directly. The per-bin structure of equation 7 lets the loss tensor $L_i(\cdot;\theta)$ over the B^M quality states be assembled from B encoded bins per modality without repeated encoder passes, and its additive projection under the product of fixed marginals yields $J_i(\cdot;\theta)$ in closed form (appendix). We penalize the smooth surrogate of $\|J_i\|_\infty$

$$\mathcal{L}_{\mathrm{QIR}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\alpha} \log \sum_{s\in\mathcal{S}} \exp\left(\alpha \, |J_i(s;\theta)|\right). \tag{8}$$

By Theorem 3, driving $\mathcal{L}_{\mathrm{QIR}}$ down shrinks the certified risk variation under every coupling with the same marginals. At evaluation we report the exact modulus Ω_i via the linear program. Since J_i is computed on the complete joint state space, QIR also captures higher-order interactions that pairwise differences would miss. It does not reward discarding modalities: the clean and branch terms below anchor each modality's predictive value, as Section 4 verifies.

Training objective. For each example i, let $\widehat{\mathcal{U}}_i$ be the admissible coupling set and $\mathcal{L}_i(\nu;\theta) = \mathbb{E}_{S\sim\nu}[\ell(F_\theta(\widetilde{C}_i(S)), y_i) + \gamma \sum_m \ell(u_{m,\theta}(\widetilde{X}_{mi}(S_m)), y_i)]$. StableFuse optimizes

$$\min_\theta \; \widehat{R}_{\mathrm{clean}}(F_\theta) + \gamma \sum_{m=1}^{M} \widehat{R}_{\mathrm{clean}}(u_{m,\theta}) + \tau \, \mathcal{L}_{\mathrm{QIR}}(\theta) + \lambda \frac{1}{N}\sum_{i=1}^{N} \max_{\nu\in\widehat{\mathcal{U}}_i} \mathcal{L}_i(\nu;\theta). \tag{9}$$

QIR is the theory-aligned term, minimizing a differentiable upper bound on the coupling-sensitivity modulus. The final term anchors the risk level on corrupted inputs, and its inner maximization selects a challenging coupling per sample while preserving every marginal. QIR instead controls the variation of this risk across couplings, the quantity bounded by Theorem 3. The clean and branch terms preserve nominal performance and modality-local evidence. The coupling construction, corruption operator, projection formula, and linear program are detailed in the appendix.

Table 1: Results under sample-local Quality CDS on four multimodal benchmarks. Values are mean $\pm$std over five seeds. Methods use the same official test split, feature backbone, and corruption couplings within each benchmark. **Bold** marks the best result and underline the second best.

	Clean			Sync. Quality		Cond. Worst QCDS		Avg. Async. QCDS	
Method	Acc.	F1	B-Acc.	Acc.	F1	Acc.	F1	Acc.	F1
AVE									
QMF (ICML'23)	84.78 $\pm$0.71	83.74 $\pm$0.60	82.90 $\pm$0.71	79.55 $\pm$0.97	77.91 $\pm$0.83	78.08 $\pm$0.79	76.22 $\pm$0.80	79.88 $\pm$0.80	78.07 $\pm$0.86
EAU (CVPR'24)	83.48 $\pm$0.99	82.31 $\pm$1.04	82.30 $\pm$1.29	77.71 $\pm$1.41	76.00 $\pm$1.66	75.73 $\pm$1.00	73.89 $\pm$1.29	77.81 $\pm$1.02	76.05 $\pm$1.23
MMPareto (ICML'24)	84.68 $\pm$1.00	83.83 $\pm$1.02	83.51 $\pm$1.10	79.13 $\pm$1.28	77.65 $\pm$1.36	77.21 $\pm$1.06	75.61 $\pm$1.10	79.38 $\pm$1.01	77.79 $\pm$1.13
I2MoE (ICML'25)	83.58 $\pm$0.61	82.19 $\pm$0.56	81.80 $\pm$0.53	78.18 $\pm$1.23	76.59 $\pm$1.11	75.95 $\pm$1.18	74.25 $\pm$1.08	78.34 $\pm$1.14	76.73 $\pm$1.12
CPSC (CVPR'26)	85.07 $\pm$0.39	84.32 $\pm$0.46	83.63 $\pm$0.41	79.64 $\pm$0.91	78.15 $\pm$0.86	78.06 $\pm$0.78	76.36 $\pm$0.70	79.75 $\pm$0.72	78.16 $\pm$0.71
StableFuse (Ours)	**85.52** $\pm$0.67	**84.61** $\pm$0.87	**83.82** $\pm$1.01	**81.12** $\pm$0.62	**79.77** $\pm$0.57	**78.89** $\pm$0.69	**77.29** $\pm$0.72	**80.89** $\pm$0.58	**79.42** $\pm$0.55
CREMA-D									
QMF (ICML'23)	61.94 $\pm$0.41	62.17 $\pm$0.41	62.29 $\pm$0.52	43.63 $\pm$0.74	43.82 $\pm$0.73	40.17 $\pm$0.40	40.30 $\pm$0.43	43.84 $\pm$0.47	44.00 $\pm$0.47
EAU (CVPR'24)	62.63 $\pm$1.31	62.81 $\pm$1.31	62.87 $\pm$1.20	42.64 $\pm$0.53	42.80 $\pm$0.52	38.94 $\pm$0.56	39.07 $\pm$0.48	42.83 $\pm$0.34	42.96 $\pm$0.30
MMPareto (ICML'24)	63.04 $\pm$1.13	63.36 $\pm$1.07	63.50 $\pm$1.19	44.08 $\pm$0.82	44.32 $\pm$0.86	38.94 $\pm$0.62	39.13 $\pm$0.66	44.08 $\pm$0.69	44.31 $\pm$0.72
I2MoE (ICML'25)	62.10 $\pm$0.48	62.38 $\pm$0.57	62.45 $\pm$0.44	43.71 $\pm$0.79	43.89 $\pm$0.87	37.91 $\pm$0.90	37.99 $\pm$0.98	43.96 $\pm$1.03	44.11 $\pm$1.13
CPSC (CVPR'26)	62.77 $\pm$0.89	63.05 $\pm$0.89	63.13 $\pm$0.87	43.43 $\pm$0.49	43.59 $\pm$0.51	39.77 $\pm$0.39	39.92 $\pm$0.41	43.69 $\pm$0.36	43.87 $\pm$0.39
StableFuse (Ours)	**63.47** $\pm$1.10	**63.61** $\pm$1.14	**63.92** $\pm$1.11	**47.54** $\pm$0.76	**47.66** $\pm$0.78	**43.47** $\pm$0.74	**43.61** $\pm$0.76	**47.72** $\pm$0.42	**47.85** $\pm$0.42
CMU-MOSI									
QMF (ICML'23)	71.98 $\pm$1.46	71.74 $\pm$1.33	72.96 $\pm$0.93	60.69 $\pm$0.58	60.44 $\pm$0.58	59.33 $\pm$0.33	59.05 $\pm$0.30	60.72 $\pm$0.41	60.46 $\pm$0.37
EAU (CVPR'24)	72.54 $\pm$0.75	72.16 $\pm$0.70	73.00 $\pm$0.58	61.73 $\pm$0.60	61.36 $\pm$0.52	58.19 $\pm$0.65	57.80 $\pm$0.62	61.57 $\pm$0.52	61.20 $\pm$0.45
MMPareto (ICML'24)	71.81 $\pm$0.85	71.48 $\pm$0.88	72.43 $\pm$0.95	61.14 $\pm$0.55	60.90 $\pm$0.51	57.86 $\pm$0.29	57.63 $\pm$0.26	60.93 $\pm$0.54	60.70 $\pm$0.51
I2MoE (ICML'25)	71.05 $\pm$1.00	70.84 $\pm$1.01	72.11 $\pm$1.05	61.10 $\pm$0.64	60.77 $\pm$0.49	53.23 $\pm$0.35	52.94 $\pm$0.18	61.36 $\pm$0.63	61.02 $\pm$0.46
CPSC (CVPR'26)	70.41 $\pm$0.64	70.15 $\pm$0.63	71.27 $\pm$0.63	60.60 $\pm$0.37	60.35 $\pm$0.38	59.01 $\pm$0.42	58.71 $\pm$0.42	60.57 $\pm$0.21	60.31 $\pm$0.20
StableFuse (Ours)	**74.14** $\pm$1.30	**73.62** $\pm$1.22	**74.19** $\pm$1.08	**62.20** $\pm$0.19	**61.71** $\pm$0.16	**60.15** $\pm$0.27	**59.65** $\pm$0.23	**62.17** $\pm$0.14	**61.64** $\pm$0.12
CH-SIMS									
QMF (ICML'23)	68.75 $\pm$1.31	56.67 $\pm$1.34	56.55 $\pm$1.16	65.39 $\pm$0.77	54.71 $\pm$0.97	63.46 $\pm$0.84	53.23 $\pm$0.79	65.42 $\pm$0.60	54.78 $\pm$0.55
EAU (CVPR'24)	65.86 $\pm$0.80	56.15 $\pm$1.25	55.76 $\pm$1.24	61.23 $\pm$0.53	52.04 $\pm$0.91	56.79 $\pm$0.65	48.34 $\pm$0.82	60.65 $\pm$0.63	51.53 $\pm$0.82
MMPareto (ICML'24)	66.91 $\pm$0.75	57.21 $\pm$0.76	56.85 $\pm$0.74	63.12 $\pm$1.09	53.49 $\pm$1.42	60.23 $\pm$0.76	51.34 $\pm$0.88	63.13 $\pm$0.65	53.70 $\pm$0.82
I2MoE (ICML'25)	67.00 $\pm$1.11	57.46 $\pm$1.87	57.05 $\pm$1.78	63.34 $\pm$1.52	53.94 $\pm$1.30	59.16 $\pm$0.69	49.95 $\pm$0.46	63.31 $\pm$1.26	53.88 $\pm$0.85
CPSC (CVPR'26)	69.02 $\pm$1.26	58.44 $\pm$1.47	58.03 $\pm$1.39	64.99 $\pm$0.68	55.04 $\pm$0.57	62.92 $\pm$0.40	53.51 $\pm$0.55	65.17 $\pm$0.50	55.05 $\pm$0.55
StableFuse (Ours)	**69.37** $\pm$1.15	**58.54** $\pm$0.93	**58.25** $\pm$0.91	**66.02** $\pm$0.73	**56.06** $\pm$0.76	**64.02** $\pm$0.54	**54.50** $\pm$0.58	**65.90** $\pm$0.69	**56.02** $\pm$0.58

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks and baselines. We evaluate StableFuse on AVE Tian et al. (2018), CREMA-D Cao et al. (2014), CMU-MOSI Zadeh et al. (2016), and CH-SIMS Yu et al. (2020), covering audio visual event recognition, emotion recognition, and multimodal sentiment analysis. We compare with QMF Zhang et al. (2023), EAU Gao et al. (2024), CPSC Jiang et al. (2026), MMPareto Wei & Hu (2024), and I2MoE Xin et al. (2025). All methods follow the official evaluation protocol and use the same modality features, backbone, label space, and temporal partition.

Quality CDS and evaluation. Each modality is divided into four temporal bins, with three consecutive bins corrupted by previous observation holding. Corruption amount, severity, and unimodal location marginals remain fixed. Synchronous quality aligns failure locations, while Quality CDS changes only their joint arrangement through complete couplings within the same sample. We report clean, synchronous, average Quality CDS, and conditional worst Quality CDS performance, where the worst coupling is selected per sample by mean prediction loss. Since all couplings preserve unimodal marginals, performance differences isolate cross modal quality dependence. We report accuracy, Macro-F1, and balanced accuracy on clean data, and accuracy and Macro-F1 under corruption, averaged over five seeds. Further details are provided in the appendix.

4.2 MAIN RESULTS

Table 1 shows that StableFuse achieves the strongest overall performance across all four benchmarks. It obtains the best clean accuracy while remaining consistently strongest under synchronous quality and average Quality CDS, indicating that robustness does not come at the cost of nominal performance. The advantage is preserved under conditional worst Quality CDS. StableFuse achieves the highest

worst case accuracy and Macro-F1 on every benchmark, with particularly clear gains on CREMA-D. Since every coupling has identical unimodal corruption marginals, these improvements reflect robustness to changes in cross modal quality dependence rather than weaker corruption.

4.3 Empirical Validation of Quality CDS Theory

Table 2: Stable modalities but unstable fusion under Quality CDS. Unimodal performance is unchanged across complete couplings, while fusion accuracy varies under fixed and conditional coupling shifts. Values are mean $\pm$std over five seeds. **Bold** marks the best result and underline the second best.

	AVE		CREMA-D		CMU-MOSI		CH-SIMS	
Method	**Fixed Gap**	**Cond. Gap**	**Fixed Gap**	**Cond. Gap**	**Fixed Gap**	**Cond. Gap**	**Fixed Gap**	**Cond. Gap**
QMF	0.86 $\pm$0.44	2.23 $\pm$0.33	0.80 $\pm$0.62	4.07 $\pm$0.81	0.92 $\pm$0.19	2.06 $\pm$0.15	1.06 $\pm$0.43	2.00 $\pm$0.30
EAU	0.53 $\pm$0.47	1.98 $\pm$0.62	0.66 $\pm$0.37	3.70 $\pm$0.41	0.92 $\pm$0.13	3.55 $\pm$0.48	1.87 $\pm$0.38	4.43 $\pm$0.60
MMPareto	0.42 $\pm$0.51	1.92 $\pm$0.32	0.87 $\pm$0.37	5.14 $\pm$0.40	1.33 $\pm$0.17	3.28 $\pm$0.29	0.93 $\pm$0.57	2.89 $\pm$0.51
I2MoE	0.40 $\pm$0.28	2.24 $\pm$0.52	0.70 $\pm$0.63	5.80 $\pm$0.19	0.90 $\pm$0.23	7.87 $\pm$0.42	**0.83** $\pm$0.41	4.18 $\pm$0.93
CPSC	0.60 $\pm$0.10	1.58 $\pm$0.44	0.47 $\pm$0.51	3.66 $\pm$0.49	0.98 $\pm$0.50	1.60 $\pm$0.67	0.85 $\pm$0.25	2.07 $\pm$0.47
StableFuse	**0.22** $\pm$0.22	**1.47** $\pm$0.32	**0.45** $\pm$0.36	**3.46** $\pm$0.37	**0.82** $\pm$0.49	**1.36** $\pm$0.40	1.06 $\pm$0.14	**1.94** $\pm$0.23

Stable modalities, unstable fusion. Table 2 directly validates the central CDS phenomenon. Although unimodal performance remains exactly unchanged across complete couplings, every method exhibits nonzero fusion gaps on all four benchmarks. Conditional gaps further exceed fixed gaps in all 24 method benchmark settings, showing that globally shared couplings can substantially underestimate CDS sensitivity.

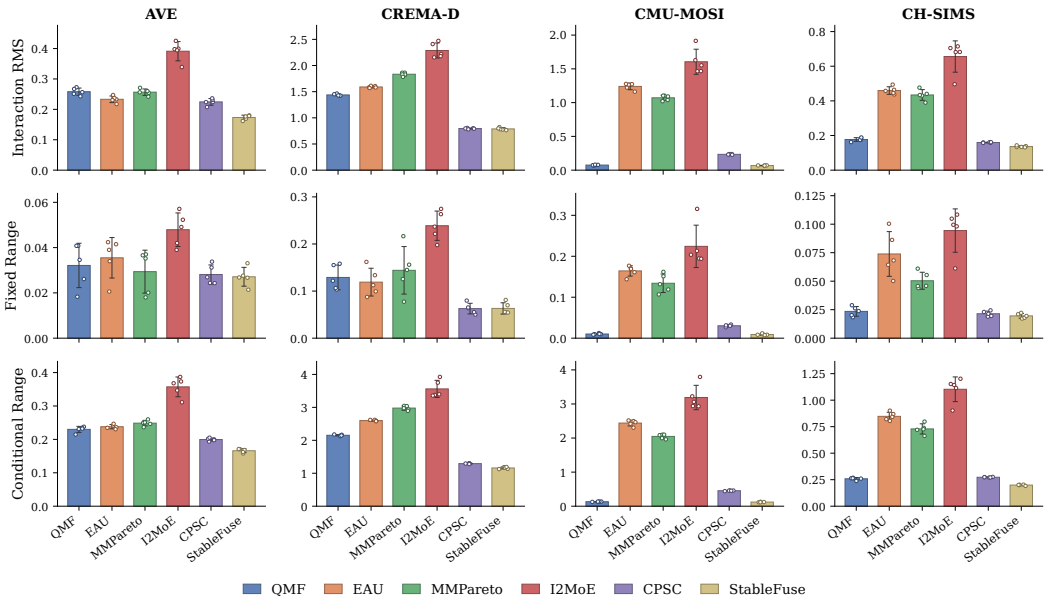

Figure 3: Empirical quality interaction and coupling sensitivity across four benchmarks. Interaction RMS measures nonadditive interaction, while Fixed and Conditional Range measure sensitivity to globally shared and sample dependent couplings with identical unimodal marginals. Bars show the mean over five seeds, error bars the standard deviation, and dots individual seeds.

Interaction and coupling sensitivity. Figure 3 connects this behavior to our theoretical characterization. Unlike the accuracy gaps of Table 2, which conflate risk level with coupling variation, these loss based ranges are the quantities aligned with Theorem 3. Across benchmarks, methods with stronger interaction generally show greater coupling sensitivity, consistent with Theorem 1, while StableFuse attains the lowest ranges. Conditional Range is also substantially larger for several interaction sensitive methods, supporting our conditional formulation of Quality CDS.

Table 3: Exact certified modulus Ω (Sec. A.8), mean $\pm$std over five seeds, for StableFuse trained with and without QIR ($\tau = 0$). Loss scales differ across benchmarks, but relative reduction is scale free.

	AVE	CREMA-D	CMU-MOSI	CH-SIMS
w/o QIR (τ=0)	0.247 $\pm$0.006	2.010 $\pm$0.014	0.140 $\pm$0.006	0.265 $\pm$0.009
StableFuse	**0.195** $\pm$0.008	**0.091** $\pm$0.006	**0.062** $\pm$0.002	**0.161** $\pm$0.007
Reduction	20.9%	95.5%	55.8%	39.4%

QIR shrinks the certified modulus. Table 3 reports the modulus Ω_i of Appendix A.8, comparing StableFuse trained with and without the QIR term ($\tau = 0$). QIR reduces the certified modulus on every benchmark, by 21% on AVE up to 95% on CREMA-D, and the reduction tracks the worst case gains of Table 1: the largest modulus reduction coincides with the largest worst case improvement. Clean accuracy (Table 1) and the fusion gain over the best unimodal branch are preserved, verifying that certified coupling sensitivity decreases without sacrificing complementary evidence.

4.4 Ablation Study

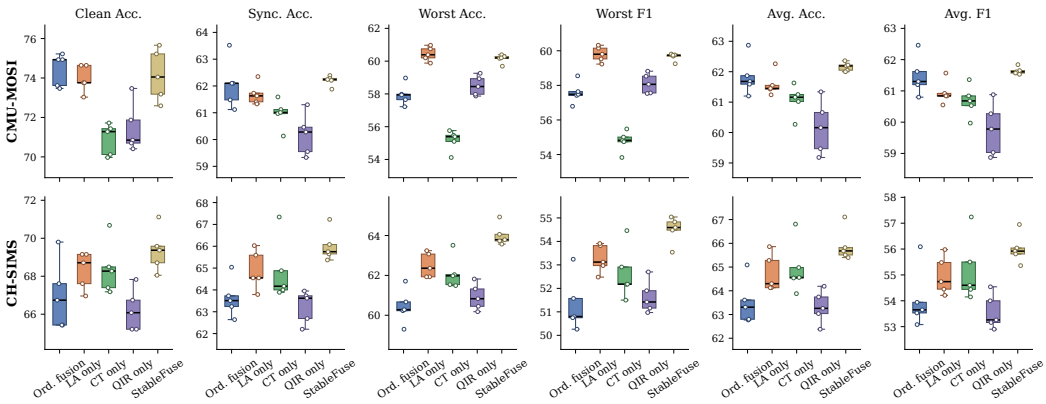

Figure 4: Ablation of StableFuse. LA, CT, and QIR denote local aggregation, coupling training, and interaction regularization. The full model performs best. Dots mark five individual seeds.

Component ablation. Figure 4 examines the three components of StableFuse. QIR alone already improves worst case performance on ordinary joint fusion, showing the regularizer is not tied to the modality local architecture, while CT anchors corrupted performance and LA preserves clean accuracy. Their combination is most consistent across clean, synchronous, average, and conditional worst settings, matching the intended division between risk level and coupling variation.

Table 4: Equal Quality CDS exposure on AVE. The baseline matches StableFuse in corruption, coupling family, robust weight, and budget. Mean$\pm$std, five seeds. Bold best, underline second best.

Variant	**Clean Acc.**	**Clean F1**	**Sync. Acc.**	**Worst Acc.**	**Worst F1**	**Avg. Acc.**
CPSC	<u>85.07</u>$\pm$0.39	<u>84.32</u>$\pm$0.46	79.64$\pm$0.91	78.06$\pm$0.78	76.36$\pm$0.70	79.75$\pm$0.72
+ equal QCDS exposure	84.23$\pm$0.52	83.44$\pm$0.56	<u>79.79</u>$\pm$0.79	<u>78.41</u>$\pm$0.92	<u>76.95</u>$\pm$0.84	<u>79.96</u>$\pm$0.79
StableFuse	**85.52**$\pm$0.67	**84.61**$\pm$0.87	**81.12**$\pm$0.62	**78.89**$\pm$0.69	**77.29**$\pm$0.72	**80.90**$\pm$0.57

Equal Quality CDS exposure. We test whether StableFuse gains simply from additional exposure to Quality CDS during training. In Table 4, CPSC is trained with the same corruption operator, coupling family, robust-training weight, and coupling candidates as StableFuse. Equal exposure improves robustness under coupled corruption but does not reproduce the worst case gains of StableFuse and slightly reduces clean accuracy: exposure alone is insufficient, and the QIR-based StableFuse design provides benefits beyond robust data augmentation.

Table 5: Conditional worst Quality CDS under mean imputation (MI) and previous observation holding (PH). Values are mean ± standard deviation over five seeds.

Dataset	Operator	Metric	QMF	EAU	MMPareto	I2MoE	CPSC	StableFuse
AVE	MI	Acc.	81.98±0.71	80.56±1.33	81.74±1.26	80.67±1.09	81.97±0.74	**83.47**±0.52
		F1	80.47±0.68	79.13±1.59	80.29±1.46	79.14±0.99	80.55±0.70	**82.31**±0.57
	PH	Acc.	80.85±0.76	79.04±1.15	80.46±1.01	79.20±0.99	80.67±0.63	**82.04**±0.74
		F1	79.23±0.81	77.43±1.34	79.00±1.16	77.65±0.98	79.16±0.60	**80.65**±0.88

4.5 Robustness across Quality Operators

Table 5 evaluates mean imputation and previous observation holding under the same Quality CDS protocol on AVE. These operators perturb local evidence differently, testing whether robustness depends on the corruption used above. StableFuse achieves the best accuracy and Macro-F1 under both operators. Since coupling construction and unimodal marginals remain unchanged, robustness to quality dependence holds across local degradations. All four benchmarks (Appendix A.1) show this.

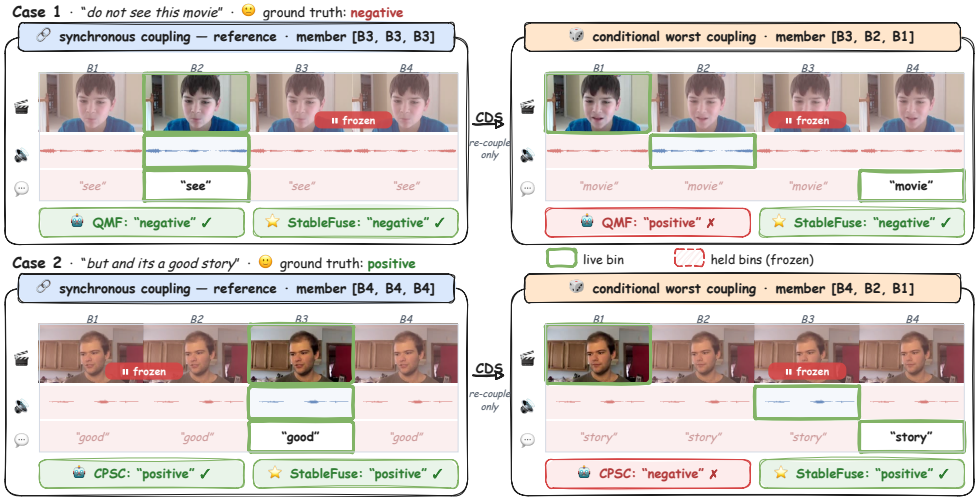

Figure 5: Qualitative Quality CDS examples. Changing only the failure coupling can flip baseline predictions, while StableFuse stays correct. Green boxes mark live bins and dashed boxes held bins.

4.6 Qualitative Analysis

Figure 5 presents two representative Quality CDS cases. For each sample, the unimodal corruption marginals remain unchanged, while only the cross modal coupling of local failures changes. Under the conditional worst coupling, QMF and CPSC change from correct to incorrect predictions even though no modality becomes individually more corrupted. StableFuse remains correct in both cases, illustrating its reduced sensitivity to unstable joint failure patterns.

5 Conclusion

We study conditional dependence shift (CDS), where individual modality quality distributions remain fixed while their cross modal dependence changes. We prove that CDS sensitivity is governed by cross modal quality interaction and introduce a coupling-sensitivity modulus certifying risk variation across fixed-marginal couplings. We therefore propose StableFuse, combining modality local aggregation with quality interaction regularization, which directly minimizes an upper bound on this modulus. Across four benchmarks, StableFuse improves average and conditional worst CDS performance while reducing the certified modulus and preserving clean accuracy. Our results highlight quality dependence as a distinct source of multimodal instability beyond individual modality degradation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with mathematical derivations, coding, and improving the clarity and readability of the manuscript. All AI-assisted mathematical derivations were independently checked by the authors, all AI-assisted code was verified and tested for correctness, and all AI-assisted text was reviewed and revised by the authors. We did not use generative AI tools to generate research hypotheses, design the methodology or experiments, or analyze and interpret experimental results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

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.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

MingCai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In *Forty-second International Conference on Machine Learning*, 2025.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Hao Dong, Eleni Chatzi, and Olga Fink. Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization. In *International Conference on Learning Representations*, volume 2025, pp. 77426–77451, 2025.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

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

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26866–26875. IEEE, 2024.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Zirun Guo and Tao Jin. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In *International Conference on Learning Representations*, volume 2025, pp. 2215–2232, 2025.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. *Advances in Neural Information Processing Systems*, 37:67446–67466, 2024.

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. *arXiv preprint arXiv:2605.03820*, 2026.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25864–25873. IEEE, 2025.

Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pp. 171–187. Springer, 2024.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Jiacheng Li and Songhe Feng. Bridging modalities via progressive re-alignment for multimodal test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22931–22939, 2026.

Zichao Li, Xueru Wen, Jie Lou, Yuqiu Ji, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models. *arXiv preprint arXiv:2503.03122*, 2025.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, 2021(DB1):1, 2021.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023.

Fengming Lin, Xiaolei Fang, and Zheming Gao. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization*, 12(1):159–212, 2022.

Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023.

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. Mitigating spurious correlations in zero-shot multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 14727–14753, 2025.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18156–18165. IEEE, 2022.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8237. IEEE, 2022.

Giovanni Puccetti and Ruodu Wang. Extremal dependence concepts. 2015.

Ludger Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14 (3):623–632, 1982.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European conference on computer vision*, pp. 252–268. Springer, 2018.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977*, 2017.

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

Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2776–2784, 2023.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. *Proceedings of machine learning research*, 267:68870, 2025.

Yihao Xue, Siddharth Joshi, Dang Nguyen, and Baharan Mirzasoleiman. Understanding the robustness of multi-modal contrastive learning to distribution shift. In *International Conference on Learning Representations*, volume 2024, pp. 19864–19896, 2024.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *International Conference on Learning Representations*, volume 2024, pp. 55036–55055, 2024.

Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *International Conference on Machine Learning*, pp. 39365–39379. PMLR, 2023.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3718–3727, 2020.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27446–27456. IEEE, 2024.

Lin Zhu, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. Bayesian cross-modal alignment learning for few-shot out-of-distribution generalization. In *Proceedings of the aaai conference on artificial intelligence*, volume 37, pp. 11461–11469, 2023.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hyper-modality enhancement for multimodal sentiment analysis with missing modalities. *Advances in Neural Information Processing Systems*, 38:142100–142129, 2026.

A Appendix

A.1 Complete Results across Quality Operators

Table 6: Conditional worst Quality CDS under mean imputation (MI) and previous observation holding (PH). Values are mean ± standard deviation over five seeds.

Dataset	Operator	Metric	QMF	EAU	MMPareto	I2MoE	CPSC	**StableFuse**
AVE	MI	Acc.	81.98±0.71	80.56±1.33	81.74±1.26	80.67±1.09	81.97±0.74	**83.47**±0.52
		F1	80.47±0.68	79.13±1.59	80.29±1.46	79.14±0.99	80.55±0.70	**82.31**±0.57
	PH	Acc.	80.85±0.76	79.04±1.15	80.46±1.01	79.20±0.99	80.67±0.63	**82.04**±0.74
		F1	79.23±0.81	77.43±1.34	79.00±1.16	77.65±0.98	79.16±0.60	**80.65**±0.88
CREMA-D	MI	Acc.	48.56±0.51	47.84±0.31	47.92±0.52	47.76±1.24	48.66±0.38	**51.44**±0.52
		F1	48.70±0.49	48.06±0.35	48.19±0.53	47.99±1.32	48.86±0.48	**51.57**±0.48
	PH	Acc.	45.48±0.46	45.10±0.49	44.75±0.69	44.14±0.87	45.40±0.44	**49.19**±0.33
		F1	45.64±0.44	45.27±0.53	44.98±0.71	44.33±0.93	45.62±0.43	**49.31**±0.38
CMU-MOSI	MI	Acc.	64.15±0.59	64.00±0.68	63.88±0.47	61.24±0.49	64.07±0.30	**65.05**±0.39
		F1	63.79±0.50	63.58±0.67	63.59±0.47	60.91±0.45	63.75±0.27	**64.45**±0.32
	PH	Acc.	61.83±0.58	61.79±0.49	61.15±0.77	58.00±0.72	61.36±0.42	**62.70**±0.38
		F1	61.57±0.50	61.39±0.43	60.87±0.74	57.71±0.62	61.09±0.42	**62.18**±0.38
CH-SIMS	MI	Acc.	66.23±0.33	61.26±0.62	63.51±0.43	62.99±0.99	66.04±0.37	**66.48**±0.77
		F1	55.77±0.38	52.36±0.57	54.40±0.59	53.66±0.95	56.23±0.51	**56.61**±0.78
	PH	Acc.	65.90±0.38	60.38±0.93	63.02±0.70	62.18±1.00	65.13±0.47	**65.91**±0.67
		F1	55.05±0.46	51.45±1.11	53.58±0.87	52.61±0.61	54.87±0.47	**55.73**±0.56

Table 6 reports the complete results under mean imputation and previous observation holding across all four benchmarks. The same conditional Quality CDS protocol is used as in the main paper, with the coupling construction and unimodal quality marginals kept fixed. StableFuse achieves the best mean accuracy and Macro-F1 across all benchmark and operator combinations, confirming that its robustness is consistent across different forms of local quality degradation.

A.2 Qualitative analysis of fused predictions under Quality CDS

Figure 6 visualizes the fused decision geometry of each method under conditional Quality CDS on CREMA-D. For every test sample we form the four members of its conditional worst coupling (selected as in Section 4.1), average the fused predictive distribution over the members, and embed these distributions with t-SNE (perplexity 50, fixed seed), colored by ground-truth label. All methods use seed 0 and the archived per-state predictions underlying the Cond. Worst columns of Table 1. Since every member preserves the same unimodal corruption marginals, panel differences reflect only how consistently each method fuses the same local evidence under different couplings. A predictor that flips decisions across members is pulled toward the center of the simplex and its class structure fragments, as seen for the baselines, whereas StableFuse keeps fused decisions stable and class regions coherent. Quantitatively, StableFuse attains the highest kNN purity (0.44 vs. 0.37–0.40) and the only positive silhouette score ($+0.015$ vs. -0.013 to -0.045), matching the quantitative results: identical marginals, yet the fused geometry degrades under re-coupled failures for interaction-reliant baselines while remaining stable for StableFuse.

A.3 Hyperparameter sensitivity

Figure 7 sweeps the two robustness weights on all four benchmarks with three seeds per configuration, fixing the other weight at its per-benchmark default and keeping all remaining hyperparameters as in Section A.13. Accuracy is stable across two orders of magnitude of the QIR weight $\tau \in [0.01, 1]$: clean, worst case, and average accuracy vary within roughly one standard deviation on every

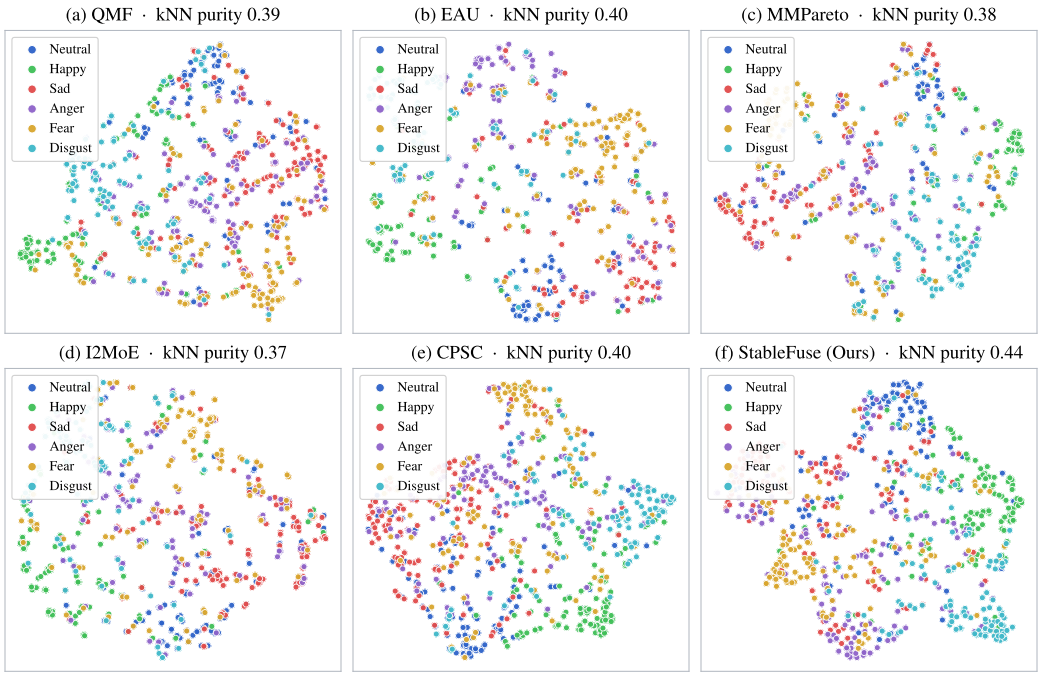

Figure 6: t-SNE of fused predictions under conditional worst Quality CDS on CREMA-D. Each point is one test sample's fused prediction averaged over its four conditional worst coupling members; colors denote ground-truth labels, and panel titles report kNN label purity (15 neighbors). All members share the same corruption marginals, so panel differences isolate coupling sensitivity. StableFuse preserves the most coherent class structure (0.44 vs. 0.37–0.40).

benchmark, so the regularizer requires no benchmark-specific tuning and we use $\tau = 0.1$ throughout. This insensitivity is consistent with the switch like behavior of QIR: a small weight already suffices to suppress the interaction residual, after which larger weights bring little further change. The CT weight λ acts as the level anchor described in Section 3.3: $\lambda = 0$, which corresponds to StableFuse without coupling training, yields the lowest worst case accuracy on every benchmark, and accuracy improves steadily with λ on AVE and CH-SIMS, whereas a large weight degrades clean accuracy on CMU-MOSI, motivating the per-benchmark choice of λ. Sensitivity remains mild around the selected values, so neither weight requires careful tuning.

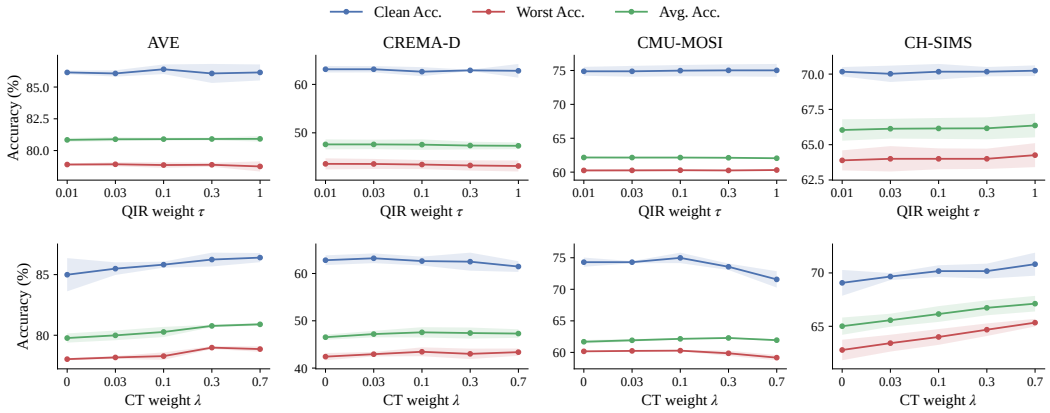

Figure 7: Sensitivity to the QIR weight τ (top) and CT weight λ (bottom): clean, conditional worst, and average accuracy, mean $\pm$ std over three seeds.

A.4 Proof of Theorem 1

Fix clean content c and label y. Let $\mu_1, \ldots, \mu_M$ denote the fixed quality marginals and define

$$\mu^{\otimes} = \prod_{m=1}^{M} \mu_m. \tag{10}$$

Recall that

$$\overline{L}_{c,y} = \mathbb{E}_{S\sim\mu^{\otimes}}\left[L_{c,y}(S)\right] \tag{11}$$

and

$$\overline{L}^{m}_{c,y}(s_m) = \mathbb{E}_{S_{-m}\sim\prod_{j\neq m}\mu_j}\left[L_{c,y}(s_m, S_{-m})\right]. \tag{12}$$

The quality interaction residual is

$$J_{c,y}(s) = L_{c,y}(s) - \sum_{m=1}^{M} \overline{L}^{m}_{c,y}(s_m) + (M-1)\overline{L}_{c,y}. \tag{13}$$

Suppose first that $J_{c,y}(s) = 0$ for every valid joint quality state. Then

$$L_{c,y}(s) = \sum_{m=1}^{M} \overline{L}^{m}_{c,y}(s_m) - (M-1)\overline{L}_{c,y}. \tag{14}$$

The loss is therefore additive in the individual modality quality states.

Let ν be any joint distribution with marginals $\mu_1, \ldots, \mu_M$. Taking expectation under ν gives

$$\mathbb{E}_{S\sim\nu}\left[L_{c,y}(S)\right] = \sum_{m=1}^{M} \mathbb{E}_{S_m\sim\mu_m}\left[\overline{L}^{m}_{c,y}(S_m)\right] - (M-1)\overline{L}_{c,y} \tag{15}$$

$$= \overline{L}_{c,y}. \tag{16}$$

The expected loss is therefore identical for every coupling with the same quality marginals.

For the converse, assume that the expected loss is invariant to every joint distribution with marginals $\mu_1, \ldots, \mu_M$. Let $\mathcal{V}$ denote the space of signed functions δ on $\mathcal{S}(c)$ whose one dimensional marginals vanish:

$$\sum_{s_{-m}} \delta(s_m, s_{-m}) = 0 \tag{17}$$

for every modality m and every state s_m.

Since every μ_m has full support, the product distribution $\mu^{\otimes}$ also has full support. For any $\delta \in \mathcal{V}$, there exists a sufficiently small $\epsilon > 0$ such that

$$\mu^{\otimes} + \epsilon\delta \qquad \text{and} \qquad \mu^{\otimes} - \epsilon\delta \tag{18}$$

are both valid probability distributions with the same marginals. By the assumed invariance,

$$\langle \delta, L_{c,y} \rangle = 0 \tag{19}$$

for every $\delta \in \mathcal{V}$.

Thus $L_{c,y}$ lies in the orthogonal complement of all perturbations that preserve the individual marginals. This orthogonal complement is exactly the space of additive functions of the form

$$L_{c,y}(s) = a_0 + \sum_{m=1}^{M} a_m(s_m). \tag{20}$$

Substituting this form into equation 13 gives

$$J_{c,y}(s) = 0 \tag{21}$$

for every valid state. This proves Theorem 1.

A.5 Proof of Theorem 2

Under CDS,

$$Q(C,Y) = P_0(C,Y). \tag{22}$$

Let $\nu^Q_{c,y}$ and $\nu^0_{c,y}$ denote the conditional joint distributions of the quality states under Q and P_0. The corresponding risks are

$$R_Q(f) = \mathbb{E}_{(C,Y)\sim P_0}\left[\sum_{s\in\mathcal{S}(C)} \nu^Q_{C,Y}(s) L_{C,Y}(s)\right], \tag{23}$$

$$R_{P_0}(f) = \mathbb{E}_{(C,Y)\sim P_0}\left[\sum_{s\in\mathcal{S}(C)} \nu^0_{C,Y}(s) L_{C,Y}(s)\right]. \tag{24}$$

Using equation 13, write

$$L_{c,y}(s) = J_{c,y}(s) + \sum_{m=1}^{M} \overline{L}^m_{c,y}(s_m) - (M-1)\overline{L}_{c,y}. \tag{25}$$

Taking the difference between the two risks gives one term involving $J_{C,Y}$ and one term involving the additive part. The latter vanishes because $\nu^Q_{C,Y}$ and $\nu^0_{C,Y}$ have the same marginal distribution for every S_m. Hence

$$\mathbb{E}_{\nu^Q_{C,Y}}\left[\overline{L}^m_{C,Y}(S_m)\right] = \mathbb{E}_{\nu^0_{C,Y}}\left[\overline{L}^m_{C,Y}(S_m)\right] \tag{26}$$

for every modality m.

The constant term $\overline{L}_{C,Y}$ also cancels because both joint distributions have total mass one. We therefore obtain

$$R_Q(f) - R_{P_0}(f) = \mathbb{E}_{(C,Y)\sim P_0}\left[\left\langle \nu^Q_{C,Y} - \nu^0_{C,Y}, J_{C,Y}\right\rangle\right]. \tag{27}$$

This proves Theorem 2.

A.6 Pairwise mixed quality interaction

The interaction residual has a particularly simple interpretation for two modalities. Let their quality states be s_1 and s_2. For any s_1, s_1' and s_2, s_2', define the rectangular mixed difference

$$\begin{aligned}\Delta L_{c,y} = {} & L_{c,y}(s_1,s_2) - L_{c,y}(s_1',s_2) \\ & - L_{c,y}(s_1,s_2') + L_{c,y}(s_1',s_2').\end{aligned} \tag{28}$$

If the loss is additive,

$$L_{c,y}(s_1,s_2) = a_{c,y}(s_1) + b_{c,y}(s_2), \tag{29}$$

then every mixed difference is zero. Conversely, suppose every mixed difference vanishes. Fix reference states s_1^0 and s_2^0 and define

$$a_{c,y}(s_1) = L_{c,y}(s_1, s_2^0) \tag{30}$$

and

$$b_{c,y}(s_2) = L_{c,y}(s_1^0, s_2) - L_{c,y}(s_1^0, s_2^0). \tag{31}$$

Setting $(s_1', s_2') = (s_1^0, s_2^0)$ in equation 28 gives

$$L_{c,y}(s_1,s_2) = a_{c,y}(s_1) + b_{c,y}(s_2). \tag{32}$$

Thus, the pairwise mixed difference measures whether the effect of one modality quality state depends on the quality state of the other modality. Nonzero mixed differences are precisely the pairwise form of the quality interaction exposed to CDS. For $M \geq 3$ modalities, vanishing pairwise differences do not rule out higher-order interaction; this is why QIR penalizes the full residual $J_{c,y}$ on the complete joint state space rather than pairwise terms.

A.7 Proof of Theorem 3

Equivalence $\Omega_{c,y}(f) = 0 \iff J_{c,y} = 0$. If $J_{c,y} = 0$, then $L_{c,y} = A_{c,y} \in \mathcal{H}_{\text{add}}$ and hence $\Omega_{c,y}(f) = 0$. Conversely, on a finite state space $\mathcal{H}_{\text{add}}$ is a finite dimensional linear subspace, hence closed, so the infimum in the definition of $\Omega_{c,y}$ is attained. If $\Omega_{c,y}(f) = 0$, the minimizer $A^\star$ satisfies $\|L_{c,y} - A^\star\|_\infty = 0$, so $L_{c,y} = A^\star \in \mathcal{H}_{\text{add}}$. Since every quality marginal has full support, the orthogonal projection in equation 3 is well defined and leaves $\mathcal{H}_{\text{add}}$ fixed, so $A_{c,y} = L_{c,y}$ and $J_{c,y} = 0$.

Coupling sensitivity bound. Let $\nu, \nu' \in \mathcal{C}_{c,y}$ and let $A \in \mathcal{H}_{\text{add}}$ be arbitrary, $A(s) = a_0 + \sum_m a_m(s_m)$. Because ν and ν' share the same modality marginals $\mu_1, \dots, \mu_M$,

$$\mathbb{E}_\nu[A] = a_0 + \sum_{m=1}^{M} \mathbb{E}_{S_m \sim \mu_m}\left[a_m(S_m)\right] = \mathbb{E}_{\nu'}[A]. \tag{33}$$

Hence

$$\mathbb{E}_\nu[L_{c,y}] - \mathbb{E}_{\nu'}[L_{c,y}] = \mathbb{E}_\nu[L_{c,y} - A] - \mathbb{E}_{\nu'}[L_{c,y} - A] \tag{34}$$

$$= \langle \nu - \nu', L_{c,y} - A \rangle . \tag{35}$$

By Hölder's inequality,

$$|\mathbb{E}_\nu[L_{c,y}] - \mathbb{E}_{\nu'}[L_{c,y}]| \le \|\nu - \nu'\|_1 \, \|L_{c,y} - A\|_\infty . \tag{36}$$

Since $A \in \mathcal{H}_{\text{add}}$ is arbitrary and the infimum is attained, taking the infimum over A yields

$$|\mathbb{E}_\nu[L_{c,y}] - \mathbb{E}_{\nu'}[L_{c,y}]| \le \|\nu - \nu'\|_1 \, \Omega_{c,y}(f). \tag{37}$$

The projection $A_{c,y}$ is one admissible choice of A, so $\Omega_{c,y}(f) \le \|J_{c,y}\|_\infty$, which gives the second inequality in equation 6.

Risk consequence. Apply the bound with $\nu = \nu^Q_{c,y}$ and $\nu' = \nu^0_{c,y}$ for every (c, y) and take the expectation over $(C, Y) \sim P_0$ using Theorem 2:

$$|R_Q(f) - R_{P_0}(f)| \le \mathbb{E}_{(C,Y)\sim P_0}\left[\|\nu^Q_{C,Y} - \nu^0_{C,Y}\|_1 \, \Omega_{C,Y}(f)\right] \le \rho \, \mathbb{E}_{(C,Y)\sim P_0}\left[\Omega_{C,Y}(f)\right]. \tag{38}$$

If $J_{c,y} = 0$ for every valid state, then $\Omega_{c,y}(f) = 0$ and the risk is invariant to every quality coupling with the same modality marginals. This proves Theorem 3.

A.8 Linear program for the coupling-sensitivity modulus

On a finite state space, $\Omega_{c,y}(f)$ is the value of a small linear program. Introduce variables $a_0 \in \mathbb{R}$, $a_m(s_m) \in \mathbb{R}$ for every modality m and state s_m, and a slack $t \ge 0$:

$$\begin{aligned} \Omega_{c,y}(f) = \min_{a_0, \{a_m\}, t} \quad & t \\ \text{s.t.} \quad & -t \le L_{c,y}(s) - a_0 - \sum_{m=1}^{M} a_m(s_m) \le t, \qquad \forall s \in \mathcal{S}. \end{aligned} \tag{39}$$

The additive parameterization has a gauge freedom (constants can be shifted between a_0 and the a_m); we remove it by constraining $\sum_{s_m} a_m(s_m) = 0$ for every m. With $B = 4$ states per modality, the program has $1 + M(B-1) + 1$ free variables and $2B^M$ inequality constraints: 8 variables and 32 constraints for two modality benchmarks, and 11 variables and 128 constraints for three modality benchmarks. Solving one instance takes a fraction of a millisecond, so the exact modulus can be computed for every test example.

Certificate evaluation protocol. For each test example i we assemble the loss tensor $L_i(\cdot; \theta)$ over all B^M joint quality states (Section A.9), solve equation 39, and report the resulting exact modulus Ω_i. We compare its distribution for StableFuse trained with and without the QIR term ($\tau = 0$) to verify that QIR shrinks the certified sensitivity. Because the loss tensor depends on the label y_i, Ω_i is an evaluation-time diagnostic rather than a deployable quantity.

Label-free variant. A label-free upper bound is obtained by maximizing over candidate labels,

$$\bar{\Omega}_c(f) = \max_{y \in \mathcal{Y}} \Omega_{c,y}(f), \tag{40}$$

which requires only the input c: the loss tensor is evaluated for every candidate y and the largest modulus is kept. Since $\bar{\Omega}_c \geq \Omega_{c,y}$ for the realized label, Theorem 3 holds with $\bar{\Omega}_C$ in place of $\Omega_{C,Y}$, at the cost of a looser bound and a factor $|\mathcal{Y}|$ in computation.

A.9 Efficient loss tensor assembly

Assembling the loss tensor over all B^M joint quality states does not require B^M encoder passes. Under the per-bin architecture of Section A.10, each bin is encoded independently as $h_{m,t} = E_m(\widetilde{x}_{m,t})$. The previous observation holding operator equation 51 replaces the content of a held bin by that of an already existing bin, so the encoding of every corrupted bin equals an already computed clean bin encoding $h_{m,p(s_m)}$. Consequently, for every quality state s_m the corrupted modality summary $z_m(s_m)$ is a fixed average of the B clean bin encodings, and the branch evidence $u_m(s_m) = H_m(z_m(s_m))$ requires only one head pass per state. The same holds for mean imputation, whose held-bin encoding is shared across states.

The full computation therefore costs MB encoder passes and MB head passes, after which every fused logit

$$F\big(\widetilde{C}_i(s)\big) = \sum_{m=1}^{M} a_m \, \mathrm{sg}\big(c_m(s_m)\big) \, u_m(s_m) \tag{41}$$

is a sum of precomputed per-modality terms, and the B^M losses $L_i(s;\theta)$ reduce to B^M cross entropy evaluations on cached logits (64 for three modalities, 16 for two). The additive projection equation 13 is a closed-form averaging of this tensor, so $J_i(\cdot;\theta)$ and the QIR penalty add negligible overhead per training step.

A.10 StableFuse parameterization

StableFuse estimates predictive evidence and confidence independently within each modality. Each modality is divided into B temporal regions. For modality m and region t, the observed feature is encoded as

$$h_{m,t} = E_m(\widetilde{x}_{m,t}). \tag{42}$$

The modality representation is obtained using temporal aggregation,

$$z_m = \frac{1}{B} \sum_{t=1}^{B} h_{m,t}. \tag{43}$$

A modality specific prediction head then produces

$$u_m = H_m(z_m). \tag{44}$$

The confidence score is computed only from the corresponding modality prediction,

$$c_m = \frac{1}{\kappa} \operatorname{logsumexp}(u_m), \tag{45}$$

where κ controls its scale. We use $\kappa = 10$ in all experiments. When the confidence is used in the fused prediction, its gradient is stopped. The separate reliability ranking objective described in Section A.13 provides supervision for the confidence estimate.

Let $\overline{a}_m > 0$ be the scale assigned to modality m. We normalize the scales according to

$$a_m = \frac{\overline{a}_m}{\frac{1}{M}\sum_{j=1}^{M} \overline{a}_j}. \tag{46}$$

The scales are fixed within each benchmark and shared across all samples. They do not depend on the corruption locations or the representations of other modalities.

The final StableFuse prediction is

$$F(\widetilde{C}) = \sum_{m=1}^{M} a_m \,\mathrm{sg}(c_m) u_m, \tag{47}$$

where sg denotes gradient stopping.

Every confidence estimate is computed before evidence from different modalities is combined. StableFuse therefore does not inspect whether two modalities are corrupted at the same temporal region. The model retains evidence from all available modalities while avoiding direct functions of cross modal corruption locations.

The additive prediction rule does not imply that the quality interaction $J_{c,y}$ is exactly zero, because the final classification loss is nonlinear. The residual interaction is instead controlled explicitly by the QIR penalty (Section A.13), whose value is certified at evaluation by the linear program of Section A.8. No explicit cross modal interaction branch is used in the reported main results.

A.11 Quality coupling construction

We construct Quality CDS entirely within each original sample. No feature, temporal region, or modality is exchanged across different examples.

Each modality stream is divided into

$$B = 4 \tag{48}$$

temporal regions. The quality state s_m specifies the starting location of a local sensor stall in modality m. The stall affects the three consecutive regions

$$\mathcal{B}(s_m) = \{s_m, s_m + 1, s_m + 2\} \pmod{B}. \tag{49}$$

The most recent reliable observation is located at

$$p(s_m) = s_m - 1 \pmod{B}. \tag{50}$$

We simulate previous observation holding by defining

$$\widetilde{x}_{m,t}^{(s_m)} = \begin{cases} x_{m,p(s_m)}, & t \in \mathcal{B}(s_m), \\ x_{m,t}, & t \notin \mathcal{B}(s_m). \end{cases} \tag{51}$$

Every modality therefore experiences a stall of the same duration and uses the same corruption operator in every environment. The corruption changes which temporal evidence remains current, while preserving the original sample identity and label. Since $B = 4$, each quality state retains one reliable temporal region and replaces the other three regions with the preceding reliable observation.

To construct a complete quality coupling, we use the first modality as the reference. Let

$$r \sim \mathrm{Uniform}\,(\{1, \ldots, B\}) \tag{52}$$

denote its stall starting location. For each modality m, define

$$S_m = \pi_m(r), \tag{53}$$

where

$$\pi_1 = \mathrm{id} \tag{54}$$

and every remaining π_m is a permutation of $\{1, \ldots, B\}$.

A coupling $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ induces

$$\nu^{\boldsymbol{\pi}} = \frac{1}{B} \sum_{r=1}^{B} \delta_{(r, \pi_2(r), \ldots, \pi_M(r))}. \tag{55}$$

Because every π_m is a permutation, each possible stall starting location appears exactly once for modality m. Hence

$$P_{\nu^{\boldsymbol{\pi}}}(S_m = t) = \frac{1}{B} \tag{56}$$

for every modality m, every location t, and every admissible coupling. Consequently, every temporal region has the same corruption probability and every modality has exactly the same quality state distribution under all couplings. The couplings differ only in how the modality quality states occur together.

The synchronous quality coupling is obtained by setting

$$\pi_m = \mathrm{id} \tag{57}$$

for every modality. All modalities then experience the stall at corresponding temporal locations. Their remaining reliable temporal evidence therefore has the largest possible overlap.

The other permutation choices produce asynchronous couplings. Under these couplings, every modality experiences the same stall duration and the same uniform distribution over stall locations. Only the cross modal arrangement of these locations changes. The overlap of reliable temporal evidence can therefore decrease without increasing the corruption applied to any individual modality.

For M modalities, the complete family contains

$$|\mathcal{U}| = (B!)^{M-1} \tag{58}$$

couplings. With $B = 4$, each two modality benchmark contains

$$4! = 24 \tag{59}$$

complete couplings. Each three modality benchmark contains

$$(4!)^2 = 576 \tag{60}$$

complete couplings.

A.12 CDS EVALUATION DETAILS

Clean evaluation uses the original multimodal observations without local corruption. It measures nominal predictive performance and is not itself part of the fixed marginal CDS comparison.

Synchronous quality evaluation uses the identity coupling. For each test sample, we evaluate all B constituent states of this complete coupling. Every modality experiences a three region stall, and the stall begins at the same temporal location across modalities.

Let $\mathcal{U}_{\mathrm{CDS}}$ denote the complete coupling family excluding the synchronous coupling. For every $\nu \in \mathcal{U}_{\mathrm{CDS}}$, we evaluate all B constituent corruption states defined by equation 55. Since every complete coupling contains the same number of states and preserves every modality quality marginal, differences between coupling results arise only from the joint arrangement of the local stalls.

For a performance metric Metric, average asynchronous Quality CDS performance is

$$\mathrm{AvgCDS} = \frac{1}{|\mathcal{U}_{\mathrm{CDS}}|} \sum_{\nu \in \mathcal{U}_{\mathrm{CDS}}} \mathrm{Metric}(\nu). \tag{61}$$

The conditional worst coupling is selected separately for each test sample. For sample i and coupling ν, define its mean state loss as

$$\bar{\ell}_i(\nu) = \frac{1}{B} \sum_{r=1}^{B} \ell\left(F\left(\widetilde{C}_i\left(S^{\nu}(r)\right)\right), y_i\right), \tag{62}$$

where $S^{\nu}(r)$ denotes the joint quality state induced by coupling ν and reference location r. The conditional worst asynchronous coupling is

$$\nu_i^{\star} = \arg\max_{\nu \in \mathcal{U}_{\mathrm{CDS}}} \bar{\ell}_i(\nu). \tag{63}$$

After selecting $\nu_i^{\star}$, we collect its B predictions for every test sample and compute accuracy and Macro-F1 over the resulting predictions. The selection is based on mean cross entropy rather than on the reported metric. It therefore yields one valid complete coupling for each sample while preserving

the complete quality marginal of every modality. It never selects individual corruption states and never exchanges content between samples.

The coupling may depend on the sample and label because the theory permits the conditional quality distribution $\nu_{c,y}$ to vary with clean content c and label y. Conditional worst evaluation is an oracle stress test of the bounded Quality CDS family, in the same sense as the certificate Ω_i of Section A.8. It is not intended to represent a corruption selection mechanism available during ordinary deployment.

All methods are evaluated using the same full official test examples, temporal partition, corruption operator, and complete coupling family. Clean evaluation reports accuracy, Macro-F1, and balanced accuracy. Synchronous quality, average asynchronous Quality CDS, and conditional worst asynchronous Quality CDS report accuracy and Macro-F1.

A.13 Training details

StableFuse is trained using supervision for both the fused prediction and the modality specific evidence. For training example i and complete coupling ν, let

$$\widehat{R}_{i,\nu}(F) = \frac{1}{B}\sum_{r=1}^{B} \ell\left(F\left(\widetilde{C}_i\left(S^{\nu}(r)\right)\right), y_i\right) \tag{64}$$

denote the fused classification risk averaged over all states of the complete coupling. We define $\widehat{R}_{i,\nu}(u_m)$ in the same way for modality m.

QIR term. For every example i in the batch, we assemble the loss tensor $L_i(\cdot;\theta)$ over all B^M joint quality states from the cached per-bin encodings (Section A.9), compute the residual $J_i(\cdot;\theta)$ by the closed-form projection equation 13, and evaluate the smooth surrogate

$$\mathcal{L}_{\mathrm{QIR}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\alpha}\log\sum_{s\in\mathcal{S}}\exp\left(\alpha\left|J_i(s;\theta)\right|\right) \tag{65}$$

of equation 8. We use $\alpha = 10$ in all experiments; the surrogate upper bounds $\|J_i\|_\infty$ and converges to it as $\alpha \to \infty$.

Coupling training term. Enumerating the complete coupling family during every optimization step is unnecessary. For every example i and update, we independently sample a set $\mathcal{K}_i \subset \widehat{\mathcal{U}}_i$ containing two complete couplings. Each candidate preserves the uniform quality marginal of every modality. We then use the candidate with the largest average classification risk. The maximum over $\mathcal{K}_i$ is a stochastic approximation to the conditional inner maximization in equation 9: the selected coupling may differ across training samples, and every candidate remains a complete coupling, so the adversarial selection never changes an individual modality quality marginal.

Implemented objective. The implemented objective is

$$\begin{aligned}\mathcal{L} = \widehat{R}_{\mathrm{clean}}(F) + \gamma\sum_{m=1}^{M}\widehat{R}_{\mathrm{clean}}(u_m) + \tau\,\mathcal{L}_{\mathrm{QIR}} \\ + \frac{\lambda}{N}\sum_{i=1}^{N}\max_{\nu\in\mathcal{K}_i}\left[\widehat{R}_{i,\nu}(F) + \gamma\sum_{m=1}^{M}\widehat{R}_{i,\nu}(u_m)\right] + \beta\mathcal{L}_{\mathrm{rel}}.\end{aligned} \tag{66}$$

Here, $\mathcal{L}_{\mathrm{rel}}$ is a modality local ranking objective that aligns the ordering of confidence scores with the ordering of accumulated unimodal prediction losses. It does not compare quality states or representations across modalities. We use $\gamma = 0.35$, $\beta = 1$, and $\tau = 0.1$ for all benchmarks.

The QIR term directly minimizes a differentiable upper bound on the coupling-sensitivity modulus certified by Theorem 3, so that the risk variation admissible under any coupling with the same modality marginals is explicitly driven down during training. The branch terms preserve predictive evidence within individual modalities under clean and corrupted observations, preventing the trivial solution in which the interaction is removed by discarding modalities. The coupling training term

anchors the risk level on corrupted inputs, complementing the variation control of QIR, and the reliability ranking term calibrates the modality local confidence used by the additive predictor.

All experiments use AdamW with an initial learning rate of 3×10^{-4}, weight decay 10^{-4}, and batch size 128. Class weights are defined as

$$w_k = \left(\frac{N}{K N_k} \right)^p , \tag{67}$$

where N_k is the number of training examples in class k, K is the number of classes, and p is selected per benchmark.

A.14 Qualitative case studies under Quality CDS

Figures 8 and 9 show eight real CMU-MOSI test segments under the sample-local Quality CDS protocol of Section 4.1. For each case we contrast two members of the same corruption process: a synchronous member, in which the live bins of all modalities occupy the same temporal slot, and one member of the sample's conditional worst coupling. The worst member shares its audio state with the synchronous member, so only the visual and textual failure locations are re-coupled. Both members apply the same corruption amount and severity and preserve every unimodal quality marginal; the panels differ only in how the live bins align across modalities.

Each panel shows the actually observed inputs: video frames at the four bin centers, the per-bin audio waveform, and the transcript split across bins, with live bins outlined and frozen bins shaded. Under the synchronous member every method predicts correctly; under the conditional worst member the baseline flips on every case (for example, "this movie isnt just bad its diabolical" flips from negative to positive under QMF), while StableFuse keeps its prediction. The eight cases cover all five baselines, both polarities, and eight distinct speakers.

These cases are illustrative rather than exhaustive: they were selected as samples on which the baseline is correct under the synchronous member but flips under the conditional worst member while StableFuse is correct under both (seed 0). Such flips are not rare: 46 test samples satisfy this criterion for QMF and 68 for CPSC. All predictions are read off the archived per-state evaluation of Table 1 and are exactly reproducible.



Figure 8: Qualitative Quality CDS cases on CMU-MOSI (cases 1–4). Left: synchronous member (live bins aligned); right: one conditional worst member (live bins misaligned). Green outlines mark live bins; shaded bins hold a frozen copy. The audio stream is identical in both panels and every unimodal marginal is preserved. The baseline flips under re-coupled failures while StableFuse remains correct.



Figure 9: Qualitative Quality CDS cases on CMU-MOSI (cases 5–8), continuing Figure 8. The eight cases cover all five baselines and both polarities.