CoRiM: Conflict-driven Risk Minimization for Dynamic Multimodal Fusion

Supplementary Material

A. Theoretical Analysis

This appendix provides the theoretical analysis supporting the CoRiM framework presented in Section 3. We first formally define the full Modality Conflict Risk (MCR) function, then analyze its non-convexity and smoothness properties, and finally provide a detailed proof of the convergence guarantees for the Frank-Wolfe algorithm when applied to this objective.

A.1. Full Formulation of Modality Conflict Risk

Our goal is to find a weight vector $w \in \Delta^{M-1}$ that minimizes a risk function $\mathcal{R}(w)$ capturing both prediction uncertainty and inter-modal inconsistency. We restate the full MCR function:

$$\mathcal{R}(w) = \underbrace{\alpha H(p_w)}_{\text{Term 1: Fused Entropy}} + \underbrace{\beta H_m(w)}_{\text{Term 2: Modal Confidence}} + \underbrace{\gamma \text{JS}_m(w)}_{\text{Term 3: JS Consistency}} \tag{1}$$

A.2. Analysis of Smoothness

While $\mathcal{R}(w)$ is non-convex, we can show it is L-smooth (i.e., its gradient is Lipschitz continuous). This is the key property that allows us to apply modern non-convex FW theory.

Lemma 1 (Smoothness of $\mathcal{R}(w)$)**.** *The risk function $\mathcal{R}(w)$ is differentiable and L-smooth on the simplex Δ^{M-1}, assuming the class probabilities $p_m(c)$ and $p_w(c)$ are bounded away from zero (e.g., $p(c) \geq \epsilon > 0$), which is a standard assumption or enforced by softmax calculation.*

Proof. A function is L-smooth if its gradient is L-Lipschitz continuous, which is equivalent to its Hessian matrix $H = \nabla^2 \mathcal{R}(w)$ having a bounded spectral norm ($L = \sup_{w \in \Delta^{M-1}} \|H(w)\|_2 < \infty$). We prove this by showing that every component of H is bounded.

The Hessian H is an $M \times M$ matrix where each element $(H)_{jk} = \frac{\partial^2 \mathcal{R}}{\partial w_j \partial w_k}$. The total Hessian is the sum of the Hessians for each term: $H = H_\alpha + H_\beta + H_\gamma$.

We analyze each term's Hessian $(H)_{jk}$. First, the Hessians for Modal Confidence Term is straightforward:

$$(H_\beta)_{jk} = \frac{\partial^2}{\partial w_j \partial w_k} \left(\beta \sum_{i=1}^{M} w_i H(p_i) \right) = 0 \tag{2}$$

$H_\beta = \mathbf{0}$ is constant and thus bounded.

Second, for Fused Entropy Term, we differentiate its gradient $\nabla_{w_k} \mathcal{R}_\alpha = -\alpha \sum_c p_k(c)[1 + \log p_w(c)]$ with respect to w_j. Since $p_k(c)$ is a constant w.r.t w_j, we have:

$$\begin{aligned}(H_\alpha)_{jk} &= \frac{\partial}{\partial w_j}(\nabla_{w_k} \mathcal{R}_\alpha) \\ &= -\alpha \sum_c p_k(c) \left(\frac{\partial}{\partial w_j} \log p_w(c) \right) \\ &= -\alpha \sum_c \frac{p_k(c) p_j(c)}{p_w(c)}\end{aligned} \tag{3}$$

Finally, for JS Consistency Term, we differentiate its gradient $\nabla_{w_k} \mathcal{R}_\gamma = \frac{\gamma}{2M} \sum_c p_k(c) (\sum_{i=1}^{M} \log \frac{p_w(c)}{m_i(c)})$, where $m_i = \frac{1}{2}(p_i + p_w)$.

$$\begin{aligned}(H_\gamma)_{jk} &= \frac{\partial}{\partial w_j}(\nabla_{w_k} \mathcal{R}_\gamma) \\ &= \frac{\gamma}{2M} \sum_c p_k(c) \left(\sum_{i=1}^{M} \left[\frac{\partial \log p_w(c)}{\partial w_j} - \frac{\partial \log m_i(c)}{\partial w_j} \right] \right)\end{aligned} \tag{4}$$

Using the same chain rules , where $\frac{\partial \log p_w(c)}{\partial w_j} = \frac{p_j(c)}{p_w(c)}$ and $\frac{\partial \log m_i(c)}{\partial w_j} = \frac{1}{2} \frac{1}{m_i(c)} p_j(c)$, we get:

$$(H_\gamma)_{jk} = \frac{\gamma}{2M} \sum_c p_k(c) p_j(c) \left(\sum_{i=1}^{M} \left[\frac{1}{p_w(c)} - \frac{1}{2m_i(c)} \right] \right) \tag{5}$$

All terms $p_m(c)$ are outputs of a softmax, and we assume they are bounded away from zero, $p_m(c) \geq \epsilon > 0$. Since $w \in \Delta^{M-1}$, $p_w(c) = \sum w_m p_m(c) \geq \epsilon$ and $m_i(c) = \frac{1}{2}(p_i(c) + p_w(c)) \geq \epsilon$. Therefore, all denominators ($1/p_w(c)$, $1/m_i(c)$) are bounded above by $1/\epsilon$. Since all probabilities $p_j(c), p_k(c)$ are bounded by 1, every element $(H)_{jk}$ of the Hessian matrix is a sum of bounded terms, and is thus bounded. A matrix with bounded elements over a compact domain Δ^{M-1} has a bounded spectral norm. So the gradient $\nabla \mathcal{R}(w)$ is Lipschitz continuous. Therefore, there exists a constant $L < \infty$ such that:

$$\|\nabla \mathcal{R}(w_1) - \nabla \mathcal{R}(w_2)\| \leq L \|w_1 - w_2\|, \quad \forall w_1, w_2 \in \Delta^{M-1} \tag{6}$$

This L-smoothness property is sufficient for guaranteeing the convergence of the FW algorithm to a stationary point. □

A.3. Convergence Analysis of Frank-Wolfe

We now provide a self-contained proof for the convergence of the FW algorithm on our non-convex but smooth objective $\mathcal{R}(w)$ over the convex set Δ^{M-1}.

Definition 1 (Frank-Wolfe Stationarity Gap)**.** *For a differentiable function $f(w)$ on a convex set $\mathcal{D}$, the FW gap (or stationarity gap) at a point w_t is defined as:*

$$G_t(w_t) := \max_{s \in \mathcal{D}} \langle \nabla f(w_t), w_t - s \rangle \tag{7}$$

A point w^* is a first-order stationary point if and only if $G_t(w^*) = 0$. Our goal is to show that $\min_t G_t(w_t) \to 0$.

Theorem 1 (FW Convergence Rate for $\mathcal{R}(w)$)**.** *Let the sequence w_t be generated by the Frank-Wolfe algorithm (Algorithm 1) applied to the L-smooth function $\mathcal{R}(w)$ on the simplex Δ^{M-1}. If the step size is chosen as $\eta_t = \eta \leq 1/\sqrt{T}$, the rate is:*

$$\min_{\{t=0,\dots,T\}} G_t(w_t) \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \tag{8}$$

If the step size is chosen as $\eta_t = \frac{2}{t+2}$, the convergence rate of the stationarity gap is:

$$\min_{\{t=0,\dots,T\}} G_t(w_t) \leq \mathcal{O}\left(\frac{1}{\log T}\right) \tag{9}$$

In either case, the algorithm is guaranteed to converge to a stationary point.

Proof. We begin from the standard descent lemma, which is guaranteed by the L-smoothness of $\mathcal{R}(w)$ (proven in Lemma 1):

$$\mathcal{R}(w_{t+1}) \leq \mathcal{R}(w_t) + \langle \nabla \mathcal{R}(w_t), w_{t+1} - w_t \rangle + \frac{L}{2} \|w_{t+1} - w_t\|^2 \tag{10}$$

The Frank-Wolfe update is $w_{t+1} = (1-\eta_t)w_t + \eta_t s_t$, where s_t is the LMO solution $s_t = \arg\min_{s \in \Delta^{M-1}} \langle s, \nabla \mathcal{R}(w_t) \rangle$. This implies $w_{t+1} - w_t = \eta_t(s_t - w_t)$. Substituting this into Eq. (10):

$$\mathcal{R}(w_{t+1}) \leq \mathcal{R}(w_t) + \eta_t \langle \nabla \mathcal{R}(w_t), s_t - w_t \rangle + \frac{L\eta_t^2}{2} \|s_t - w_t\|^2 \tag{11}$$

By the definition of the stationarity gap $G_t(w_t)$ (Definition 1) and the LMO step s_t, we know that $\langle \nabla \mathcal{R}(w_t), s_t - w_t \rangle = -G_t(w_t)$. This simplifies the inequality to:

$$\mathcal{R}(w_{t+1}) \leq \mathcal{R}(w_t) - \eta_t G_t(w_t) + \frac{L\eta_t^2}{2} \|s_t - w_t\|^2 \tag{12}$$

The domain Δ^{M-1} is compact. Let $D^2 = \sup_{w_1, w_2 \in \Delta^{M-1}} \|w_1 - w_2\|^2$ be the squared diameter of the simplex (where $D^2 \leq 2$ for $M \geq 2$). We can thus bound the last term:

$$\mathcal{R}(w_{t+1}) \leq \mathcal{R}(w_t) - \eta_t G_t(w_t) + \frac{LD^2\eta_t^2}{2} \tag{13}$$

Let $\mathcal{R}^* = \min_w \mathcal{R}(w)$ be the minimum (which exists as $\mathcal{R}$ is continuous on a compact set). Rearranging Eq. (13) to isolate the gap term, we get:

$$\eta_t G_t(w_t) \leq \mathcal{R}(w_t) - \mathcal{R}(w_{t+1}) + \frac{LD^2\eta_t^2}{2} \tag{14}$$

We now sum this inequality from $t = 0$ to $T-1$:

$$\sum_{t=0}^{T-1} \eta_t G_t(w_t) \leq \sum_{t=0}^{T-1} (\mathcal{R}(w_t) - \mathcal{R}(w_{t+1})) + \frac{LD^2}{2} \sum_{t=0}^{T-1} \eta_t^2 \tag{15}$$

The first term on the right is a telescoping series, $\sum_{t=0}^{T-1} (\mathcal{R}(w_t) - \mathcal{R}(w_{t+1})) = \mathcal{R}(w_0) - \mathcal{R}(w_T)$. Since $\mathcal{R}(w_T) \geq \mathcal{R}^*$, this sum is bounded by $C_0 = \mathcal{R}(w_0) - \mathcal{R}^*$, a constant. This yields:

$$\sum_{t=0}^{T-1} \eta_t G_t(w_t) \leq C_0 + \frac{LD^2}{2} \sum_{t=0}^{T-1} \eta_t^2 \tag{16}$$

Let $G_{\min}^{(T)} = \min_{t=0..T-1} G_t(w_t)$. We can bound the left side of Eq. (16) by $G_{\min}^{(T)} \left(\sum_{t=0}^{T-1} \eta_t\right)$:

$$G_{\min}^{(T)} \left(\sum_{t=0}^{T-1} \eta_t\right) \leq \sum_{t=0}^{T-1} \eta_t G_t(w_t) \leq C_0 + \frac{LD^2}{2} \sum_{t=0}^{T-1} \eta_t^2 \tag{17}$$

This directly implies the final rate:

$$G_{\min}^{(T)} \leq \frac{C_0 + (LD^2/2) \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} \tag{18}$$

From here, we analyze the two standard step size policies.

Case 1 (Standard Non-Convex Rate): If we choose a constant step size $\eta_t = \eta \leq 1/\sqrt{T}$, then $\sum_{t=0}^{T-1} \eta_t = T\eta \leq \sqrt{T}$ and $\sum_{t=0}^{T-1} \eta_t^2 = T\eta^2 \leq 1$. Plugging into Eq. (18):

$$G_{\min}^{(T)} \leq \frac{C_0 + LD^2/2}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \tag{19}$$

Case 2 (Step size $\eta_t = 2/(t+2)$): As $T \to \infty$, the sum $\sum_{t=0}^{T-1} \eta_t \approx 2\log(T)$ and the sum $\sum_{t=0}^{T-1} \eta_t^2$ converges to a constant. This gives:

$$G_{\min}^{(T)} \leq \frac{C_0 + (LD^2/2)C_1}{2\log(T)} = \mathcal{O}\left(\frac{1}{\log T}\right) \tag{20}$$

Both choices of step size guarantee that the stationarity gap $G_{\min}^{(T)} \to 0$ as $T \to \infty$.

This proves convergence. □

Dataset	Method	$\epsilon = 0.0$	$\epsilon = 5.0$	$\epsilon = 10.0$
MVSA	TEXT	75.61 ± 0.53	69.50 ± 1.50	47.41 ± 0.79
	IMG	64.12 ± 1.23	49.36 ± 2.02	45.00 ± 2.63
	CONCAT	65.59 ± 1.33	50.70 ± 2.65	46.12 ± 2.44
	Late Fusion	76.88 ± 1.30	63.46 ± 3.46	55.16 ± 3.60
	QMF	78.07 ± 1.10	73.85 ± 1.42	61.28 ± 2.12
	TMC	74.87 ± 2.24	66.72 ± 4.55	60.35 ± 2.79
	DynMM	79.07 ± 0.53	67.96 ± 1.65	59.21 ± 1.41
	PDF	79.94 ± 0.95	74.4 ± 1.51	63.09 ± 1.33
	Ours	**81.12** ± 0.43	**75.65**±0.70	**65.34**±1.03
UPMC FOOD 101	TEXT	86.46 ± 0.05	67.38± 0.19	43.88 ± 0.32
	IMG	64.62 ± 0.40	34.72 ± 0.53	33.03 ± 0.37
	CONCAT	88.20 ± 0.34	61.10 ± 2.02	49.86 ± 2.05
	Late Fusion	90.69 ± 0.12	68.49 ± 3.37	57.99 ± 1.59
	QMF	92.92 ± 0.11	76.03 ± 0.70	62.21 ± 0.25
	TMC	89.86 ± 0.07	73.93 ± 0.34	61.37 ± 0.21
	DynMM	92.59 ± 0.07	74.74 ± 0.19	59.68 ± 0.20
	PDF	93.32 ± 0.22	76.47 ± 0.31	62.83 ± 0.31
	Ours	**93.62**±0.15	**78.09**±0.23	**63.71**±0.18
NYUD v2	RGB	63.30 ± 0.48	53.12 ± 1.52	45.46 ± 2.07
	Depth	62.65 ± 1.22	50.95 ± 3.38	44.13 ± 3.80
	concat	69.88 ± 0.52	63.82 ± 1.46	60.03 ± 2.63
	Late Fusion	70.03 ± 0.84	64.37 ± 0.80	60.55 ± 1.65
	TMC	70.40 ± 0.31	59.33± 2.19	50.61 ± 2.87
	QMF	69.54 ± 1.06	64.10 ±1.42	60.18±1.23
	DynMM	65.50 ± 0.37	54.31 ± 1.72	46.79 ± 1.09
	PDF	71.37 ± 0.76	65.72 ± 1.72	62.56 ± 1.84
	Ours	**72.36** ± 0.45	**66.30**±0.83	**63.04** ± 0.96
SUN RGB-D	RGB#	56.78 ± 0.19	48.40 ± 1.11	42.94 ± 1.63
	Depth#	52.99 ± 0.88	37.81 ±1.14	33.07 ± 1.81
	CONCAT#	62.48± 0.50	53.30 ± 0.39	48.01 ± 0.96
	Late Fusion#	62.00 ± 0.15	52.52 ± 0.67	47.48 ± 1.40
	TMC#	60.68 ± 0.24	51.24 ± 0.96	45.66 ± 2.06
	QMF#	62.09 ± 0.56	53.40 ±0.89	45.58 ± 0.82
	PDF*	60.60 ± 0.27	51.45 ± 1.04	46.09 ± 3.08
	Ours	**63.59** ± 1.10	**55.61** ± 0.95	**48.12** ± 1.13

Table 1. **We add Gaussian noise on** 50% **modalities and** ϵ **presents the noise degree.** it shows the average and the standard deviation of classification accuracies with our method and the compared methods on four datasets. The method marked with * was replicated by us, marked with # was copied from [45], while the rest of the data is sourced from [4].

B. Implementation details.

All experiments were conducted on an RTX 4090 GPU, and all our algorithms were implemented using PyTorch. We employed Adam as the model optimizer with a weight decay of 0.1 and dropout of 0.1. Consistent with baseline implementation standards, batch_size was set to 16 on MVSA, Food101, NYUD v2, and SUN RGB-D datasets. The learning rate is 1e-4 with a warmup rate of 0.1. We adopt the early stop strategy based on validation accuracy.

For MOSI, MOSEI, and SIMS datasets, we maintained consistency with the baseline by using a batch_size of 64 and a learning rate of 0.01. More detailed hyperparameters are provided in Section E (Hyperparameter Analysis).

C. Full Results with Standard Deviation

Table 1 and Table 2 provide a full performance comparison of our CoRiM method against current SOTA methods, including PDF(ICML'24)[4] and QMF(ICML'23)[45], on four benchmark datasets under two types of noise: Gaussian and Salt-pepper. These results clearly demonstrate three

DATASET	METHOD	ϵ = 0.0	ϵ = 5.0	ϵ = 10.0
MVSA	TEXT	75.61±0.53	69.50±1.50	47.41±0.79
	IMG	64.12±1.23	56.72±1.92	50.71±3.20
	CONCAT	65.59±1.33	58.69±2.25	51.16±2.99
	LATE FUSION	76.88±1.30	67.88±1.87	55.43±1.94
	QMF	78.07±1.10	73.90±1.89	60.41±2.63
	TMC	74.87±2.24	68.02±3.07	56.62±3.67
	DYNMM	79.07±0.53	71.35±0.97	59.96±1.31
	PDF	79.94±0.95	75.11±1.15	61.97±1.14
	OURS	**81.73** ± 0.43	**77.70**±0.70	**65.85**±1.03
UPMC FOOD101	TEXT	86.44±0.02	67.41±0.20	43.89±0.33
	IMG	64.53±0.47	50.75±0.44	36.83±0.92
	CONCAT	88.22±0.36	72.49±0.75	52.10±0.97
	LATE FUSION	90.66±0.16	77.99±0.54	58.75±0.99
	QMF	92.90±0.13	80.87±0.40	61.60±0.20
	TMC	89.86±0.07	77.86±0.41	60.22± 0.43
	DYNMM	92.59±0.07	78.91±0.20	57.64±0.30
	PDF	93.32±0.22	**81.21**±0.34	61.76±0.33
	OURS	**93.62**±0.51	80.21±0.88	**62.49**±0.69
NYUD V2	RGB	62.61±1.21	49.14±1.40	34.76±1.59
	DEPTH	63.32±0.50	50.99±1.41	38.56±2.16
	CONCAT	69.88±0.52	61.41±1.69	51.65±2.94
	LATE FUSION	70.03±0.84	62.05±1.17	51.50±1.81
	TMC	70.40±0.31	59.33±1.47	45.32±2.84
	QMF	69.54±1.06	62.02±1.47	51.87±0.91
	DYNMM	65.50±0.37	52.26±1.45	38.17±1.17
	PDF	71.73±0.76	64.27±1.36	53.62±2.15
	OURS	**72.36**±0.45	**65.81**±1.19	**55.03**±1.65
SUN RGB-D	RGB#	52.99±0.88	40.42±0.99	28.15±1.00
	DEPTH#	56.78±0.19	46.36±0.82	35.66±1.44
	CONCAT#	62.48±0.50	51.09±1.91	38.61±3.07
	LATE FUSION#	62.00±0.15	51.54±2.12	39.35±2.89
	TMC#	60.68±0.24	50.88±1.28	39.61±2.30
	QMF#	62.09±0.56	52.49±1.81	40.53±2.79
	PDF*	60.60±0.27	53.22±0.76	43.30±1.84
	OURS	**63.59**±1.10	**54.14**±0.87	**45.91**±1.63

Table 2. **We add Salt-pepper noise on** 50% **modalities and** ϵ **presents the noise degree.** it shows the average and the standard deviation of classification accuracies with our method and the compared methods on four datasets. The method marked with * was replicated by us, marked with # was copied from [45], while the rest of the data is sourced from [4].

consistent conclusions: (1) On all clean data ($\epsilon = 0.0$), our method achieved the best performance. (2) As the noise intensity increased to $\epsilon = 5.0$ and $\epsilon = 10.0$, although the performance of all models degraded, our method consistently maintained the highest accuracy, demonstrating the strongest robustness. (3) Most critically, as the noise intensified, the performance gap between our method and the runner-up baseline (such as PDF) significantly widened, for example, on the MVSA dataset with Salt-pepper noise (Table 2), the performance gap at $\epsilon = 10.0$ expanded to nearly 3.9 percentage points. This strongly proves that the robust design of the CoRiM framework is far superior to existing baseline methods when handling high conflict and OOD noise.

D. Ablation Study on More Datasets

To validate the unique contributions of each component in our proposed risk function $\mathcal{R}(w)$, we conducted a comprehensive ablation study on the NYUD v2 dataset. The CoRiM model was trained by selectively removing one or more risk terms (FU: α, MC: β, JS-C: γ), and its performance was evaluated under no-noise ($\epsilon = 0.0$) and noisy ($\epsilon = 5.0, \epsilon = 10.0$) conditions. The results in Table 3 clearly demonstrate the contribution of each component. The full model (bottom row) performed best across all noise levels, proving that all three components are indispensable for achieving optimal robustness. Models using only a single component (rows 1-3) showed significant performance

FU	MC	JS-C	ϵ = 0		ϵ = 5		ϵ = 10	
			AVG	WORST	AVG	WORST	AVG	WORST
√			69.31	68.65	62.65	61.47	54.18	53.53
	√		70.64	68.96	62.54	59.80	56.13	52.60
		√	69.93	69.11	61.67	61.12	59.80	59.39
√	√		69.80	69.42	61.93	59.19	58.95	57.86
	√	√	71.08	70.86	63.32	61.09	61.15	60.33
√		√	70.36	70.26	63.82	62.72	60.25	59.74
√	√	√	**72.36**	**72.02**	**66.30**	**65.47**	**63.04**	**61.75**

Table 3. **Ablation study on NYUD v2.** To verify the effectiveness of Fused Uncertainty (FU) Term , Modal Confidence (MC), and JS Consistency (JSC) as well as the complete model.

degradation, confirming that no single risk term is sufficient to provide effective protection. Among all dual-component combinations, (MC + JS-C) (row 5) formed the strongest baseline. However, adding the Fused Uncertainty term (FU, α) to this combination (row 5-7) yielded the most significant performance leap, especially under high-noise conditions. This strongly validates that penalizing the final fused uncertainty ($\alpha H(p_w)$) can help in enhancing the model's noise-resistance performance.

E. Effect of Hyperparameters

To investigate the sensitivity of the CoRiM framework to the three key hyperparameters in the risk function $\mathcal{R}(w)$: α (Fused Uncertainty), β (Modal Confidence), and γ (JS Consistency). We conducted a series of experiments on the MVSA and NYUD v2 datasets, with the results shown in Table 4. On both datasets, the vast majority of parameter combinations achieved similar and excellent performance, which indicates that the CoRiM framework possesses high robustness. The robustness of the CoRiM framework is not heavily dependent on a specific combination of good parameters, but rather stems from the soundness of its theoretical design. As long as we ensure the model's core objective (minimizing fused uncertainty, i.e., a high α) is guaranteed, the model can operate stably under a wide range of parameter settings.

F. Case Study

To visually demonstrate the superiority of the CoRiM framework under high conflict and OOD noise, we selected some representative samples from the MVSA dataset and compared the prediction results of our method with baseline methods, as shown in Fig. 1. In the samples from rows 1-3, the Text modality conveys a clear Negative sentiment, while the Image modality presents a neutral or slightly positive sentiment. The baseline methods were misled by this conflict and incorrectly outputted a "Positive" prediction. In contrast, our CoRiM framework correctly identified this

MVSA			
α, β, γ	$\epsilon = 0.0$	$\epsilon = 5.0$	$\epsilon = 10.0$
0.2, 0.5, 0.3	79.19	73.15	66.92
0.3, 0.5, 0.5	79.96	73.73	**67.37**
0.5, 0.7, 0.3	78.80	73.47	62.82
0.5, 0.7, 0.5	78.99	73.98	62.75
0.7, 0.9, 0.7	80.92	73.54	62.94
0.7, 0.9, 0.9	79.19	74.05	63.58
0.9, 1.0, 0.9	80.15	74.05	64.16
0.9, 1.0, 1.0	79.96	73.41	63.39
1.0, 1.0, 1.0	79.58	74.25	64.10
1.0, 0.5, 0.5	**81.12**	**75.56**	65.34
NYUD v2			
α, β, γ	$\epsilon = 0.0$	$\epsilon = 5.0$	$\epsilon = 10.0$
0.2, 0.5, 0.3	70.49	62.08	61.8
0.3, 0.5, 0.5	71.41	65.44	61.02
0.5, 0.7, 0.3	70.18	62.84	61.96
0.5, 0.7, 0.5	71.56	64.99	60.27
0.7, 0.9, 0.7	71.10	65.21	59.48
0.7, 0.9, 0.9	71.25	64.53	60.88
0.9, 1.0, 0.9	70.18	63.30	61.73
0.9, 1.0, 1.0	71.71	65.14	62.30
1.0, 1.0, 1.0	70.95	64.83	60.65
0.9, 0.5, 0.9	**72.36**	**66.30**	**63.04**

Table 4. **Sensitivity analysis on MVSA and NYUD v2 datasets.**

conflict and successfully suppressed the interfering modality through risk minimization, making a "Negative" prediction consistent with the text. Furthermore, when encountering OOD noise, our risk function can detect the significant inconsistency (conflict) between the noisy modality and the clean text modality. It automatically and in real-time reduces the trust in the image modality, thereby maintaining the correct prediction result across all noise levels.

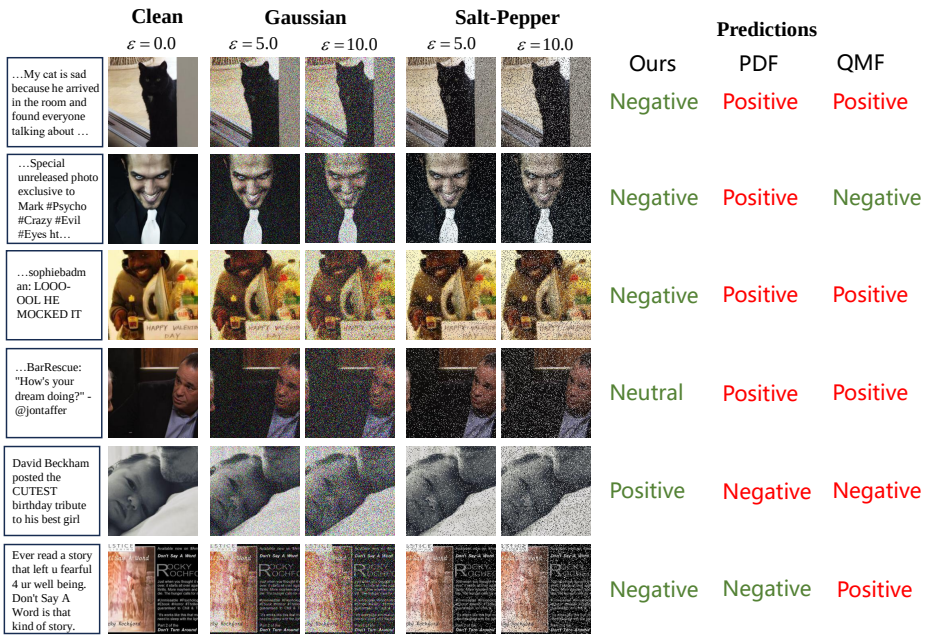

Figure 1. **Case Study on MVSA Dataset.**