APPENDIX A
PROOF OF MULTIMODAL INFORMATION BOTTLENECK LOSS.

From Eq. 2, we can obtain the loss functions of view 1 and view 2.

$$\begin{aligned}\mathcal{L}_1(\theta;\beta_1) &= I(V_1;Z_1|V_2) - \frac{1}{\beta_1}I(V_2;Z_1),\\ \mathcal{L}_2(\psi,\beta_2) &= I(V_2;Z_2|V_1) - \frac{1}{\beta_2}I(V_1;Z_2),\end{aligned} \tag{17}$$

where θ and ψ denote two functions that map V_1 and V_2 to Z_1 and Z_2, respectively. Considering the joint optimization of the two mapping networks, the total loss function can be obtained by averaging $\mathcal{L}_1(\theta;\beta_1)$ and $\mathcal{L}_2(\psi,\beta_2)$:

$$\begin{aligned}\mathcal{L}(\theta,\psi,\beta_1,\beta_2) = &\frac{I(V_1;Z_1|V_2)+I(V_2;Z_2|V_1)}{2}\\ &-\frac{\frac{1}{\beta_1}I(V_2;Z_1)+\frac{1}{\beta_2}I(V_1;Z_2)}{2}.\end{aligned} \tag{18}$$

The term $I(V_1;Z_1|V_2)$ has an upper bound:

$$\begin{aligned}I_\theta(V_1;Z_1|V_2) &= \mathbb{E}_{V_1,V_2\sim p(V_1,V_2)}\mathbb{E}_{z\sim p_\theta(Z_1|V_1)}\left[\log\frac{p_\theta(Z_1=\boldsymbol{z}|V_1=\boldsymbol{v_1})}{p_\theta(Z_1=\boldsymbol{z}|V_2=\boldsymbol{v_2})}\right]\\ &= \mathbb{E}_{V_1,V_2\sim p(V_1,V_2)}\mathbb{E}_{z\sim p_\theta(Z_1|V_1)}\left[\log\frac{p_\theta(Z_1=\boldsymbol{z}|V_1=\boldsymbol{v_1})}{p_\psi(Z_2=\boldsymbol{z}|V_2=\boldsymbol{v_2})}\cdot\frac{p_\psi(Z_2=\boldsymbol{z}|V_2=\boldsymbol{v_2})}{p_\theta(Z_1=\boldsymbol{z}|V_2=\boldsymbol{v_2})}\right]\\ &= D_{KL}(p_\theta(Z_1|V_1)\|p_\psi(Z_2|V_2)) - D_{KL}(p_\theta(Z_1|V_2)\|p_\psi(Z_2|V_2))\\ &\le D_{KL}(p_\theta(Z_1|V_1)\|p_\psi(Z_2|V_2)).\end{aligned} \tag{19}$$

By analogy, $I_\theta(V_2;Z_2|V_1)$ has an upper bound of $D_{KL}(p_\theta(Z_2|V_2)\|p_\psi(Z_1|V_1))$. Thus, the upper bound of the first term of Eq. 18 can be formulated by:

$$\begin{aligned}&\frac{I(V_1;Z_1|V_2)+I(V_2;Z_2|V_1)}{2}\\ &\le \frac{1}{2}D_{KL}\left(p_\theta(Z_1|V_1)\|p_\psi(Z_2|V_2)\right)+\frac{1}{2}D_{KL}\left(p_\theta(Z_2|V_2)\|p_\psi(Z_1|V_1)\right)\\ &= D_{SKL}\left(p_\theta(Z_1|V_1)\|p_\psi(Z_2|V_2)\right),\end{aligned} \tag{20}$$

it is worth noting that the bound is tight if $p_\psi(Z_2|V_2)$ matches $p_\theta(Z_1|V_2)$. This occurs when Z_1 and Z_2 generate a consistent encoding.

Moreover, according to the chain rule of mutual information *i.e.* $I(y;z) = I(x,y;z) - I(x;z|y)$ (1) and Z_2 is the representation of V_2 *i.e.* $I_{\theta\psi}(Z_1;Z_2|V_2)=0$ (2), $I(V_2;Z_1)$ can be reformed as:

$$\begin{aligned}I_\theta(Z_1;V_2) &\overset{(1)}{=} I_{\theta\psi}(Z_1;Z_2V_2) - I_{\theta\psi}(Z_1;Z_2|V_2)\\ &\overset{(2)}{=} I_{\theta\psi}(Z_1;Z_2V_2)\\ &= I_{\theta\psi}(Z_1;Z_2) + I_{\theta\psi}(Z_1;V_2|Z_2)\\ &\ge I_{\theta\psi}(Z_1;Z_2),\end{aligned} \tag{21}$$

the bound in this equation is tight if Z_2 is sufficient for Z_1 *i.e.*, $I_{\theta\psi}(Z_1;V_2|Z_2)=0$. This occurs when Z_2 captures all the information about Z_1 (and consequently about V_1). By analogy, we can obtain that, $I_\theta(Z_2;V_1) \ge I_{\theta\psi}(Z_1;Z_2)$. Therefore, the lower bound of the second term in Eq. 18 is:

$$\frac{\frac{1}{\beta_1}I(V_2;Z_1)+\frac{1}{\beta_2}I(V_1;Z_2)}{2} \ge \frac{(\beta_1+\beta_2)}{2\beta_1\beta_2}I(Z_1;Z_2). \tag{22}$$

Based on Eq. 20 and Eq. 22, the upper bound of the original loss function (Eq. 18) is:

$$\mathcal{L}(\theta,\psi,\beta_1,\beta_2) \le D_{SKL}\left(p_\theta(Z_1|V_1)\|p_\psi(Z_2|V_2)\right) - \frac{(\beta_1+\beta_2)}{2\beta_1\beta_2}I(Z_1;Z_2). \tag{23}$$

We use the single hyperparameter $\beta = \frac{2\beta_1\beta_2}{(\beta_1+\beta_2)}$ to reorganize Eq. 23 to obtain the loss function for the multimodal information bottleneck module as follows:

$$\mathcal{L}_{MIB}(\theta,\psi,\beta) = -I(Z_1;Z_2) + \beta D_{SKL}\left(p_\theta(Z_1|V_1)\|p_\psi(Z_2|V_2)\right). \tag{24}$$

APPENDIX B
PROOF OF LEMMA 3.1

Based on the consistency condition of counterfactuals ($X = x \Longrightarrow Y_x = Y$), we can obtain:

$$x \Longrightarrow (y_x = y), \quad x' \Longrightarrow (y'_{x'} = y), \tag{25}$$

therefore, starting from the definition of PNS (Eq. 5), we can obtain:

$$\begin{aligned} y_x \wedge y_{x'} &= (y_x \wedge y'_{x'}) \wedge (x \vee x') \\ &= (y \wedge x \wedge y'_{x'}) \vee (y_x \wedge y' \wedge x'), \end{aligned} \tag{26}$$

since x and x' do not intersect, we can directly take the probability on both sides:

$$\begin{aligned} P(y_x, y'_{x'}) &= P(y'_{x'}, x, y) + P(y_x, x', y') \\ &= P(y'_{x'} \mid x, y)P(x, y) + P(y_x \mid x', y')P(x', y'). \end{aligned} \tag{27}$$

APPENDIX C
ALGORITHM PIPELINE

The algorithmic pipeline is illustrated in Algorithm 1.

Algorithm 1: **U**ncertain **C**ross-**M**odal **I**nformation **B**ottleneck and **P**robability of **N**ecessity and **S**ufficiency (**UCMIB-PNS**) guided MSA model

Input : $D = \{(\mathbb{M}_t, \mathbb{M}_a, \mathbb{M}_v), Y\}$, hyperparameter: learning rate, training epochs, batch size, codebook size N, balance coefficient β, λ, η_1, η_2, τ

Output: Prediction $\hat{y}$ # *sentiment score*

for *each epoch* **do**
  for *mini-batch* $\{(I_t, I_a, I_v),\ Y\}$ *from* D **do**
    # Unimodal feature extraction (refer to Section III-C1)
    $F_a, F_v, F_t \leftarrow Encode(I_a^i, I_v^i, I_t^i)$
    # Multimodal Fusion (refer to Section III-C2)
    $F_{ta}, F_{tv} \leftarrow Fusion(t, a), Fusion(t, v)$
    if *training* **then**
      # UCMIB module (refer to Section III-C3)
      $\mu_{tm}, \sigma_{tm} \leftarrow MLP(F_{tm})$ *# distributed representation*
      $\tilde{F}_{tm}^q, \tilde{F}_{tm}^k \leftarrow Sample(\mu_{tm}, \sigma_{tm})$ *# reparameterization*
      $\tilde{F}_{tm}^n \leftarrow Sample(codebook)$*# reparameterization*
      $\mathcal{L}_{UCMIB} \leftarrow \tilde{F}_{tm}^q, \tilde{F}_{tm}^k, \tilde{F}_{tm}^n, \beta$*# UCMIB Loss*
      $codebook \leftarrow Dequeue(codebook)$
      $codebook \leftarrow Enqueue(codebook, \mu_{tm}, \sigma_{tm})$ *# update codebook*
      # PNS estimator module (refer to Section III-C4)
      $\hat{y}_{ta}, \hat{y}_{tv}, \hat{y} \leftarrow MLP(F_{ta}, F_{tv})$*# prediction estimation*
      $P(y'_{x'}, x, y), P(y_x, x', y') \leftarrow Logit(\hat{y}_{ta}, \hat{y}_{tv}, \hat{y}, \hat{y}^*, y)$
      $W_{tm}^{PNS} \leftarrow sum(P(y'_{x'}, x, y), P(y_x, x', y'), \lambda)$ *# PNS calculation*
      $\hat{y}_{pns} \leftarrow MLP([*W_{tm}^{PNS} F_{tm}])$ *# PNS prediction*
      $\mathcal{L}_{PNS} \leftarrow loss(\hat{y}_{tm}, \hat{y}_{pns}, y)$ *# PNS loss*
      # Loss function and optimization (refer to Section III-C5)
      $\mathcal{L} \leftarrow sum(\mathcal{L}_{task}(\hat{y}, y), \mathcal{L}_{UCMIB}, \mathcal{L}_{PNS}, \eta_1, \eta_2)$ *# total loss*
      Mini-batch gradient descent
      Update model parameters
    else
      $\hat{y} \leftarrow MLP([F_{ta}; F_{tv}])$ *# prediction*

APPENDIX D
COMPREHENSIVE EXPERIMENTS OF UCMIB-PNS ON NOISE DATASETS.

We conduct comprehensive comparative experiments on the noisy CMU-MOSI and CMU-MOSEI datasets, including experiments on different modalities of noise, different types of noise, and different levels of noise intensity, to demonstrate the advantages of UCMIB-PNS under noisy conditions. The noise settings follow [36]. We report the F1-score under the non-negative/negative (NN/N) setting (on the left side of "/") and the positive/negative (P/N) setting (marked on the right side of the slash). The noise intensity ranges from 1 to 50, determined by the noisy modality type. To ensure a fair comparison, we conduct 10 experiments using different random seeds and take the average as the reported accuracy.

A. Audio Noise

Audio noise is categorized into two types: Gaussian noise and Salt-Pepper noise. The first type, Gaussian noise, is characterized by a normal distribution, meaning that the noise values are randomly distributed around a mean value with a specific standard deviation. This type of noise is commonly found in natural environments and is often used to simulate real-world acoustic disturbances. The second type, Salt-Pepper noise, is a type of impulse noise that manifests as random occurrences of very high or very low values superimposed on the audio signal. This noise is typically caused by sudden spikes or drops in signal transmission. The results on the noisy audio modality datasets are shown in Table IX.

The results in Table IX reveal that the impact of audio noise on model performance varies across different noise types, intensities, datasets, and models. For both Salt-Pepper and Gaussian noise, most models exhibit a general decline in performance as the noise intensity increases, with Salt-Pepper noise causing a more pronounced drop compared to Gaussian noise. However, some models show a slight improvement at low noise levels, suggesting that moderate noise may act as a form of data augmentation, helping the model learn more robust features [53]. The proposed UCMIB-PNS model consistently outperforms all baselines across both noise types and all noise levels, demonstrating strong robustness. UCMIB-PNS has consistently maintained a high level of accuracy under different noise conditions, while other models, such as TETFN, exhibit a sharp decline in performance at higher noise levels, highlighting their sensitivity to severe noise conditions.

B. Visual Noise

Similar to audio, visual noise is primarily categorized into two types: Gaussian noise and Salt-Pepper noise. Gaussian noise is characterized by random variations following a normal distribution, typically representing random fluctuations in pixel values. Salt-Pepper noise, on the other hand, is an impulse noise where random pixels in an image are altered to extreme values, resembling the appearance of salt and pepper sprinkled on the image. These types of noise are commonly used to simulate real-world imperfections in visual data, such as sensor malfunctions or transmission errors. The comparative experimental results on the visual noise datasets are shown in Table X.

The results in Table X indicate that visual noise, like audio noise, has a significant impact on model performance, with the effect varying across noise types, intensities, and models. For both Salt-Pepper and Gaussian noise, most models show a general decline in performance as the noise intensity increases. Salt-Pepper noise causes a more abrupt drop in performance compared to Gaussian noise, likely due to its abrupt and extreme pixel alterations. Interestingly, we also find some models exhibit a slight improvement at low noise levels. This further confirms the conclusions we obtained from the noisy dataset, as described in Appendix D-A. The UCMIB-PNS model consistently outperforms all baselines across both noise types and all noise levels, demonstrating strong robustness. Overall, the results suggest that visual noise, particularly Salt-Pepper noise, poses a significant challenge to models, but UCMIB-PNS remains highly effective in maintaining performance through an information filtering mechanism, the sufficiency and necessity of modal data are dynamically reduced when the modal quality decreases.

C. Text Noise

The text noise in our experiments includes blank noise, where tokens in the text are replaced with a mask token (e.g., [MASK] in the case of BERT-based models) at a probability of $\epsilon\%$. This type of noise simulates scenarios where parts of the text are missing or obscured, which is a common issue in real-world applications such as speech-to-text transcription errors, incomplete data, or intentional data masking for privacy purposes. The comparative experiments on the noisy datasets are shown in Table XI.

The results in Table XI show that text noise has a notable impact on model performance, with the effect varying across noise intensities and models. As the noise intensity increases, most models exhibit a decline in performance, particularly at higher noise levels (e.g., ϵ=50), where a large proportion of text tokens are masked. However, some models also show a slight improvement at low noise levels (e.g., ϵ=5), suggesting that even in the most important text modality, appropriate noise also helps the learning of sentimental representation, possibly by leveraging contextual information [87]. The UCMIB-PNS model consistently outperforms all baselines across all noise levels, demonstrating strong robustness to text noise. Overall, the results suggest that text noise, particularly at high intensities, significantly challenges models, but UCMIB-PNS remains highly effective in maintaining performance by reducing useless information and evaluating the sufficiency and necessity of sentimental cues.

APPENDIX E
COMPREHENSIVE ABLATION EXPERIMENTS OF UCMIB-PNS ON NOISE DATASETS.

We conduct ablation studies on the CMU-MOSI dataset, where we report the F1-score under the non-negative/negative

Table IX The comparison results between UCMIB-PNS and other baseline models on noisy CMU-MOSI and CMU-MOSEI under different audio noise conditions. The marks are consistent with those used in Table I.

Noisy CMU-MOSI									
Model	Clean	Salt-Pepper Noise				Gaussian Noise			
	ϵ=0	ϵ=1	ϵ=2	ϵ=5	ϵ=10	ϵ=1	ϵ=2	ϵ=5	ϵ=10
Self-MM [26]	82.68/84.91	82.75/84.46	82.33/84.33	82.18/84.17	82.09/83.45	83.24/85.61	82.97/85.00	82.50/84.04	81.76/83.27
MMIM [24]	84.00/85.98	83.45/85.84	83.19/85.38	82.29/84.60	81.98/84.12	83.85/85.92	83.84/85.75	83.70/85.43	82.30/84.14
CENET [17]	83.49/85.22	82.95/85.30	83.82/85.26	82.75/84.45	82.50/84.03	83.05/84.46	82.77/84.80	82.55/84.41	81.38/83.34
TETFN [12]	83.83/86.07	83.02/84.64	82.98/84.07	81.80/83.15	80.70/82.47	82.75/85.10	83.19/84.91	82.63/84.79	82.02/84.32
UCMIB-PNS	**85.15/87.59**	**84.75/86.38**	**84.38/86.16**	**84.27/86.03**	**84.00/86.08**	**84.25/86.35**	**84.68/85.85**	**83.74/85.81**	**83.70/85.61**
Noisy CMU-MOSEI									
Model	Clean	Salt-Pepper Noise				Gaussian Noise			
	ϵ=0	ϵ=1	ϵ=2	ϵ=5	ϵ=10	ϵ=1	ϵ=2	ϵ=5	ϵ=10
Self-MM [26]	82.95/84.93	83.14/84.86	83.08/85.26	81.40/84.77	81.07/84.49	81.98/85.52	82.07/85.46	80.90/85.13	79.00/84.32
MMIM [24]	82.66/85.94	84.58/85.42	82.52/84.31	81.92/84.28	80.45/84.03	84.94/85.28	83.50/84.56	83.36/84.08	81.51/83.96
CENET [17]	83.85/86.32	83.10/85.51	82.50/85.41	81.72/85.25	81.38/84.87	83.65/85.95	83.09/85.78	81.77/85.47	80.31/85.22
TETFN [12]	84.18/85.27	84.33/85.31	83.90/85.71	82.13/85.33	80.22/84.93	83.94/85.76	83.89/85.21	80.59/85.03	80.44/84.94
UCMIB-PNS	**85.72/87.13**	**86.29/86.75**	**85.84/86.89**	**85.38/86.68**	**85.34/86.26**	**85.62/86.44**	**85.32/86.97**	**85.37/86.32**	**85.19/86.03**

Table X The comparison results between UCMIB-PNS and other baseline models on noisy CMU-MOSI and CMU-MOSEI under different visual noise conditions. The marks are consistent with those used in Table I.

Noisy CMU-MOSI									
Model	Clean	Salt-Pepper Noise				Gaussian Noise			
	ϵ=0	ϵ=1	ϵ=2	ϵ=5	ϵ=10	ϵ=1	ϵ=2	ϵ=5	ϵ=10
Self-MM [26]	82.68/84.91	82.57/84.74	82.29/84.13	81.90/83.88	81.49/83.29	82.70/85.19	83.00/85.03	82.39/83.76	80.91/82.53
MMIM [24]	84.00/85.98	83.90/85.66	83.28/85.48	83.24/85.13	81.94/83.30	83.50/85.39	83.32/85.05	82.95/84.83	82.51/84.05
CENET [17]	83.49/85.22	83.70/84.67	82.84/84.86	82.33/84.17	81.95/84.09	82.51/84.70	82.59/84.78	82.40/84.41	82.14/83.98
TETFN [12]	83.83/86.07	83.46/85.51	82.76/84.62	81.86/84.02	81.03/82.66	83.53/84.96	82.92/84.78	82.47/84.32	81.62/83.76
UCMIB-PNS	**85.15/87.59**	**84.73/87.02**	**84.17/85.94**	**83.82/85.89**	**83.59/85.65**	**84.60/86.38**	**84.59/86.69**	**84.15/86.24**	**83.74/85.98**
Noisy CMU-MOSEI									
Model	Clean	Salt-Pepper Noise				Gaussian Noise			
	ϵ=0	ϵ=1	ϵ=2	ϵ=5	ϵ=10	ϵ=1	ϵ=2	ϵ=5	ϵ=10
Self-MM [26]	82.95/84.93	82.54/84.75	80.29/84.21	78.40/83.61	77.64/83.01	83.86/84.98	82.77/85.72	81.58/84.56	80.24/84.13
MMIM [24]	82.66/85.94	83.13/85.74	83.56/84.79	83.31/84.11	81.94/83.30	84.15/85.65	83.53/85.57	81.74/85.38	81.29/83.64
CENET [17]	83.85/86.32	83.16/85.49	83.01/85.21	81.16/84.57	80.87/84.29	83.88/85.58	84.29/85.41	82.90/85.39	80.56/85.19
TETFN [12]	84.18/85.27	83.86/85.32	84.44/84.58	80.22/84.84	78.50/83.99	83.25/85.87	81.40/85.57	80.87/85.13	77.50/83.62
UCMIB-PNS	**85.72/87.13**	**85.50/86.80**	**85.66/86.58**	**85.32/86.18**	**85.37/85.61**	**85.65/87.02**	**85.26/86.79**	**85.00/86.03**	**84.88/85.97**

(NN/N) setting (on the left side of "/") and the positive/negative (P/N) setting (on the right side of the slash). The noise settings are consistent with those defined in [36].

The effect of the Cross-Modal Information Bottleneck (UCMIB) module under noisy conditions (the top part of Table XII). We observe that disabling any sub-component of UCMIB—such as the codebook (w/o CodeBook), the uncertainty-aware contrastive learning (w/o UACL), the full contrastive learning objective (w/o $\mathcal{L}_{CL}$), or the symmetric KL divergence (w/o $\mathcal{L}_{D_{SKL}}$)—leads to notable drops in both accuracy and F1-score, especially when noise intensity increases. Among these, the removal of the entire UCMIB loss (w/o $\mathcal{L}_{UCMIB}$) causes the most substantial degradation, with performance falling below all other variants across all visual noise settings, highlighting the critical role of UCMIB in suppressing cross-modal redundancy and enhancing noise resilience. More specifically, the removal of the uncertainty-aware contrastive loss (w/o UACL) consistently results in a larger performance decline than removing the contrastive loss alone (w/o $\mathcal{L}_{CL}$), indicating that modeling the uncertainty across modalities is more crucial than the alignment mechanism itself under noisy conditions. The codebook also plays an important role in maintaining a robust representation space, especially for challenging samples affected by noise. Besides, the symmetric KL divergence contributes to stabilizing the learning of complementary features across views, further improving robustness. These findings demonstrate that UCMIB not only enhances clean-data performance but also significantly strengthens the model's ability to resist modality-specific perturbations—particularly in the visual channel, which tends to be more vulnerable to structured noise.

The effect of the PNS estimator module under noisy conditions (the bottom part of Table XII). To assess the impact of the PNS estimator on robustness against modality-specific noise, we ablate the sufficiency and necessity estimations on the text-audio branch (w/o $\mathcal{L}_{PNS_{ta}}$), the text-vision branch (w/o $\mathcal{L}_{PNS_{tv}}$), and the entire module (w/o $\mathcal{L}_{PNS}$). Across all perturbation settings, the full model with the PNS estimator achieves the best performance, while removing the entire estimator results in the most significant decline, especially under visual Gaussian noise and audio salt-and-pepper noise. This indicates that the PNS module enhances the model's ability to dynamically adjust fusion weights and rely more heavily on reliable modalities, which is essential when

Table XI The comparison results between UCMIB-PNS and other baseline models on noisy CMU-MOSI and CMU-MOSEI under different text noise conditions. The marks are consistent with those used in Table I.

Noisy CMU-MOSI					
Model	Clean	Blank Noise			
	ϵ=0	ϵ=5	ϵ=10	ϵ=20	ϵ=50
Self-MM [26]	82.68/84.91	83.09/85.37	81.88/83.70	81.19/82.84	80.91/82.53
MMIM [24]	84.00/85.98	83.25/85.95	82.10/83.94	82.07/83.90	81.54/83.51
CENET [17]	83.49/85.22	83.20/84.93	82.84/84.86	82.41/84.76	82.31/83.53
TETFN [12]	83.83/86.07	82.77/84.32	82.26/84.10	81.78/83.93	80.80/82.91
UCMIB-PNS	**85.15/87.59**	**84.51/86.61**	**84.34/86.45**	**84.51/86.14**	**83.31/84.88**

Noisy CMU-MOSEI					
Model	Clean	Blank Noise			
	ϵ=0	ϵ=5	ϵ=10	ϵ=20	ϵ=50
Self-MM [26]	82.95/84.93	83.06/85.70	82.17/85.39	81.78/84.46	81.99/83.25
MMIM [24]	82.66/85.94	82.65/84.78	83.61/84.10	82.24/83.66	80.35/81.25
CENET [17]	83.85/86.32	83.41/85.68	82.80/85.88	82.42/85.26	81.38/84.87
TETFN [12]	84.18/85.27	82.66/85.89	82.76/85.31	81.08/85.23	80.82/85.08
UCMIB-PNS	**85.72/87.13**	**85.56/86.94**	**86.15/86.37**	**85.82/85.74**	**84.17/85.12**

Table XII Ablation results on the noisy CMU-MOSI dataset under different visual perturbation conditions. The marks are consistent with those used in Table I.

Method	Audio				Visual				Text
	Salt-pepper		Gaussian		Salt-pepper		Gaussian		Blank
	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=10
UCMIB-PNS	**84.27/86.03**	**84.00/86.08**	**83.74/85.81**	**83.70/85.61**	83.82/85.89	**85.37/85.65**	**84.15/86.24**	**83.74/85.98**	**84.34/86.45**
w/o CodeBook	83.66/85.24	83.11/84.99	83.47/85.05	82.92/84.78	83.61/84.97	82.96/84.76	84.11/85.66	83.37/85.39	82.82/85.00
w/o UACL	82.70/84.41	81.95/83.78	83.21/85.09	82.54/84.56	82.59/85.06	82.45/84.53	82.41/84.81	82.40/84.04	82.77/84.97
w/o $\mathcal{L}_{CL}$	83.71/84.83	82.15/83.67	83.33/85.06	82.76/84.78	83.51/85.44	83.38/85.31	83.62/85.61	83.28/85.18	83.83/84.81
w/o $\mathcal{L}_{D_{SKL}}$	83.66/85.39	83.50/84.61	**83.74**/85.34	82.83/84.38	**84.01**/85.40	82.97/84.92	83.77/85.68	83.06/85.04	83.70/85.92
w/o $\mathcal{L}_{UCMIB}$	82.57/83.96	81.46/83.11	82.68/84.22	81.64/82.98	82.69/84.70	81.77/84.40	82.10/84.53	81.92/83.85	81.61/83.42
w/o $\mathcal{L}_{PNS_{tv}}$	83.78/85.66	83.37/84.78	83.59/85.71	83.44/85.12	83.39/**86.05**	83.27/85.54	83.16/85.58	82.52/85.44	83.94/84.92
w/o $\mathcal{L}_{PNS_{ta}}$	83.49/85.56	82.94/84.34	83.58/85.66	83.52/84.94	83.80/85.91	83.69/85.05	84.08/85.44	83.53/85.08	83.84/85.27
w/o $\mathcal{L}_{PNS}$	82.45/84.47	81.24/83.72	82.75/84.45	81.37/83.81	82.75/84.47	81.99/83.78	82.19/84.03	82.06/83.45	82.09/84.56

dealing with corrupted or uninformative signals. Interestingly, we also observe that removing a specific branch of the PNS estimator consistently weakens the model's robustness against noise in that very modality. Specifically, removing $\mathcal{L}_{PNS_{tv}}$ degrades performance more noticeably under visual noise, while removing $\mathcal{L}_{PNS_{ta}}$ leads to greater drops under audio-related perturbations. This highlights that each branch of the PNS module contributes directly to enhancing the model's ability to handle noise within its corresponding modalities. Moreover, even under extreme conditions like text blanking or strong visual noise, the complete PNS module enables the model to retain competitive performance, showcasing its capacity to emphasize modality combinations that are sufficient and necessary for accurate prediction. These observations confirm that sufficiency and necessity estimation are not only beneficial under clean settings but become especially critical for robust multimodal fusion in the presence of unpredictable or adversarial noise.

APPENDIX F
ADDITIONAL SENSITIVITY ANALYSIS OF UCMIB-PNS ON CLEAN DATASETS

In this section, following the same experimental setting as Section IV-G, we present the sensitivity analysis results for hyperparameters α and ϕ on clean datasets.

Table XIII presents the sensitivity analysis of the hyperparameter α, which controls the size of the codebook

Table XIII The sensitivity analysis of α

α	CMU-MOSI	CMU-MOSEI	CH-SIMS	UR-FUNNY
1	47.81	53.57	43.98	71.33
2	**49.12**	54.54	**44.42**	71.53
3	46.50	**55.83**	43.10	**72.13**
4	45.19	54.63	40.48	70.22

responsible for maintaining negative samples in uncertainty-aware contrastive learning. The results show that performance improves initially with larger α, but degrades beyond a certain point. This suggests that while a moderate number of negative samples enhances the model's ability to distinguish uncertain representations, excessively increasing the sample size may introduce noisy or less informative negatives, diluting the contrastive signal. Specifically, $\alpha = 2$ achieves the best performance on CMU-MOSI and CH-SIMS, while $\alpha = 3$ is optimal for CMU-MOSEI and UR-FUNNY. These findings indicate that an appropriate balance between sample diversity and relevance is critical to for uncertainty-aware contrastive learning, and that the optimal setting of α varies across datasets.

Table XIV presents the sensitivity analysis of the hyperparameter ϕ, which controls the decision threshold. The results indicate that model performance is highly sensitive to the choice of ϕ, and the optimal value differs across datasets. For CMU-MOSI and CMU-MOSEI, the best performance is achieved when $\phi = 1$, suggesting that a more tolerant thresh-

Table XIV The sensitivity analysis of ϕ.

ϕ	CMU-MOSI	CMU-MOSEI	CH-SIMS
0.1	41.98	52.29	43.33
0.3	43.73	54.04	**44.42**
0.5	46.06	54.99	42.01
1	**49.12**	**55.83**	41.35
1.5	46.21	53.72	40.04

old better suits the characteristics of these datasets. In contrast, CH-SIMS achieves its highest accuracy when $\phi = 0.3$, likely due to its narrower value range and more compact distribution. These observations highlight the importance of aligning ϕ with the properties of each dataset.

APPENDIX G
COMPREHENSIVE SENSITIVITY ANALYSIS OF UCMIB-PNS ON NOISE DATASETS

In this section, we conduct a sensitivity analysis of five key hyper-parameters in the proposed UCMIB-PNS framework: η_1, η_2, α, λ, and ϕ. For the noisy CMU-MOSI dataset, we report the F1-score under the non-negative / negative (NN/N) setting (left of "/") and the positive / negative (P/N) setting (right of "/"). The noise settings are consistent with those defined in [36].

The sensitivity results of η_1 are shown in Table XV. We observe that setting η_1 between 0.1 and 0.2 generally leads to optimal performance across most modalities and noise types, while further increasing η_1 to 0.4 often results in a decline. This suggests that a moderate level of cross-modal redundancy compression is beneficial under noisy conditions, but excessive compression may hurt performance by discarding useful inter-modal cues. In contrast to prior expectations, the audio modality demonstrates relatively stable performance across different η_1 values, with only a slight drop at $\eta_1 = 0.4$. The visual modality shows more noticeable fluctuations, especially under Gaussian noise, where overly high η_1 can significantly degrade results. The text modality remains the most robust, with performance gradually improving and peaking at $\eta_1 = 0.2$. Overall, the results highlight that the optimal setting of η_1 is modality- and noise-dependent, and that introducing an appropriate information bottleneck is particularly helpful for enhancing robustness in noisy environments.

The sensitivity results of η_2 are shown in Table XVI. Overall, model performance improves as η_2 increases from 0.0 to 0.2, with most modalities achieving their best results at $\eta_2 = 0.2$. Further increasing η_2 to 0.4 typically leads to performance degradation, suggesting that moderate strength is beneficial, while excessive strength may suppress essential modality-specific information. Among the three modalities, audio exhibits the most consistent improvement as η_2 increases from 0 to 0.2, achieving peak performance at $\eta_2 = 0.2$ and only a slight drop at $\eta_2 = 0.4$, indicating that the audio stream contains more redundancy that can be effectively compressed. Visual performance is relatively more sensitive to η_2 and shows early gains at $\eta_2 = 0.1$ or 0.2, but in some settings (e.g., Gaussian noise), performance already starts to degrade at 0.2, implying a narrower optimal range and limited redundancy in this modality. Text remains the most stable across all η_2 values, showing steady improvement up to 0.2 and only slight degradation at 0.4. These results suggest that while appropriate strength is generally beneficial in noisy environments, the optimal strength should be carefully chosen based on the redundancy and reliability of each modality.

Table XVII presents the sensitivity analysis of the hyperparameter α, which controls the size of the codebook in UCMIB. The results show that the overall performance is relatively robust to variations in α, with optimal values typically appearing at $\alpha = 2$ or $\alpha = 3$ across most settings. In the audio and visual modalities, increasing α from 1 to 2 generally improves performance, especially under high noise levels (e.g., $\epsilon = 10$), suggesting that a moderately larger pool of negative samples enhances the model's ability to learn robust representations in noisy conditions. However, performance gains plateau or slightly decline beyond $\alpha = 3$, indicating diminishing returns from further enlarging the negative set. In contrast, the text modality achieves its best performance at $\alpha = 1$, with performance decreasing steadily as α increases. This trend implies that, due to the compact and structured nature of language embeddings, introducing too many negatives may introduce noise into the contrastive objective and hinder learning. In the visual modality, a larger negative sample codebook with $\alpha = 4$ sometimes achieves the best results. Overall, the results suggest that a moderate codebook size (e.g., $\alpha = 2$ or $\alpha = 3$) is generally sufficient for audio and visual robustness, while a smaller size is preferable for textual features.

Table XVIII presents the sensitivity analysis of the hyperparameter λ on the noisy CMU-MOSI dataset. Overall, the performance exhibits an inverted U-shaped trend, with values of λ around 0.8 to 1.2 achieving the best balance between sufficiency and necessity. Setting λ to 0, which removes the necessity term, or increasing it excessively to 1.6, leads to noticeable performance degradation. Within the Audio and Visual modalities, the optimal λ shifts depending on the noise intensity: under mild noise ($\epsilon = 5$), the best results occur at $\lambda = 0.8$, while under stronger noise ($\epsilon = 10$), a higher λ value of 1.2 yields better performance. This indicates that increasing the emphasis on necessity is beneficial in noisier conditions, helping the model to discard less essential information. In contrast, the Text modality shows a more stable performance across a wide range of λ values, remaining nearly flat from 0 to 1.2 and only peaking at 1.6 in the Blank noise setting, suggesting that the textual input is already highly sufficient and can tolerate a larger necessity weight without loss of effectiveness.

Based on the sensitivity analysis results shown in Table XIX, most configurations achieve their best performance when ϕ is between 0.5 and 1.0, although some reach their peak at 0.3 or at 1.0, indicating that there is no single optimal value suitable for all cases. Under heavier noise conditions, the optimal ϕ tends to increase slightly, suggesting the need for a looser threshold to avoid overly strict judgment on corrupted but still useful predictions. The audio modality shows the most pronounced shift, peaking at $\phi = 1.0$ under light noise and at $\phi = 0.5$ under strong noise, highlighting its sensitivity to the balance between strict and lenient evaluation. The visual modality exhibits only minor variation, while the text

Table XV The sensitivity analysis of η_1 on noisy CMU-MOSI.

η_1	Audio				Visual				Text
	Salt-pepper		Gaussian		Salt-pepper		Gaussian		Blank
	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=10
0.0	82.45/84.47	81.24/83.72	82.75/84.45	81.37/83.81	82.75/84.47	81.99/83.78	82.19/84.03	82.06/83.45	82.09/84.56
0.1	**84.27/86.03**	83.06/84.31	**83.74/85.81**	**83.70/85.61**	**83.82/85.89**	82.14/83.82	84.06/86.15	**83.74/85.98**	83.85/84.83
0.2	84.13/86.04	**84.00/86.08**	83.34/85.39	82.97/85.16	83.60/85.72	**83.59/85.65**	**84.15/86.24**	82.55/84.57	**84.34/86.45**
0.4	82.98/85.01	82.79/84.67	82.79/84.83	82.68/83.91	82.51/84.37	81.67/83.98	83.16/85.19	81.28/82.60	82.97/85.16

Table XVI The sensitivity analysis of η_2 on noisy CMU-MOSI.

η_2	Audio				Visual				Text
	Salt-pepper		Gaussian		Salt-pepper		Gaussian		Blank
	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=10
0.0	82.57/83.96	81.46/83.11	82.68/84.22	81.64/82.98	82.69/84.70	81.77/84.53	82.10/84.53	81.92/83.85	81.61/83.42
0.1	83.07/85.27	82.88/84.92	83.63/85.54	82.01/83.84	**83.82/85.89**	82.84/84.86	**84.15/86.24**	83.42/85.16	82.90/85.26
0.2	**84.27/86.03**	**84.00/86.08**	**83.74/85.81**	**83.70/85.61**	83.29/85.34	**83.59/85.65**	83.89/86.12	**83.74/85.98**	**84.34/86.45**
0.4	83.22/85.10	82.86/84.58	82.96/84.68	82.09/84.24	82.83/84.70	82.52/84.85	83.08/84.33	81.63/83.44	82.79/84.34

Table XVII The sensitivity analysis of α on noisy CMU-MOSI.

α	Audio				Visual				Text
	Salt-pepper		Gaussian		Salt-pepper		Gaussian		Blank
	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=10
1	83.50/85.55	83.53/85.10	83.59/85.49	82.81/85.52	82.98/84.34	82.40/83.47	83.35/85.57	81.82/83.49	**84.34/86.45**
2	**84.27/86.03**	**84.00/86.08**	83.72/85.72	**83.70/85.61**	83.52/84.78	82.94/84.03	83.66/85.88	82.35/84.68	84.07/86.16
3	84.23/85.69	83.86/85.61	**83.74/85.81**	82.81/85.50	**83.82/85.89**	83.50/85.54	84.00/86.24	**83.74/85.98**	83.24/85.13
4	83.64/85.71	83.12/85.00	83.30/85.34	82.65/85.18	83.81/85.88	**83.59/85.65**	**84.15/86.24**	82.48/84.49	82.76/84.95

Table XVIII The sensitivity analysis of λ on noisy CMU-MOSI.

λ	Audio				Visual				Text
	Salt-pepper		Gaussian		Salt-pepper		Gaussian		Blank
	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=10
0.0	82.64/83.72	81.13/83.24	82.43/84.77	81.84/83.83	83.38/84.91	81.27/83.08	82.87/84.89	81.86/83.84	81.02/82.80
0.4	84.07/85.52	82.44/84.96	**83.74/85.81**	**83.70/85.61**	83.74/85.34	82.49/84.03	83.88/85.79	82.51/84.05	82.15/83.67
0.8	**84.27/86.03**	83.84/85.44	83.26/85.81	83.13/84.38	**83.82/85.89**	82.98/84.70	**84.15/86.24**	**83.74/85.98**	82.42/84.94
1.2	84.23/86.04	**84.00/86.08**	83.28/85.33	82.23/83.75	83.70/85.69	**83.59/85.65**	84.03/86.10	82.49/84.03	83.57/85.48
1.6	82.75/84.77	81.43/83.08	83.22/84.47	81.18/83.46	83.37/84.78	82.42/84.94	83.02/85.21	81.51/83.50	**84.34/86.45**

modality remains relatively stable across the 0.5 to 1.0 range, demonstrating its robustness to changes in this parameter.

APPENDIX H
VISUALIZATION OF EXPERIMENTAL RESULTS ON NOISY DATASET

Fig. 8 compares the noise robustness of different models across five noisy datasets: audio with Gaussian noise, visual data with Gaussian noise, blank text noise, audio with salt-and-pepper noise, and visual data with salt-and-pepper noise—based on F1 scores under the positive/negative setting.

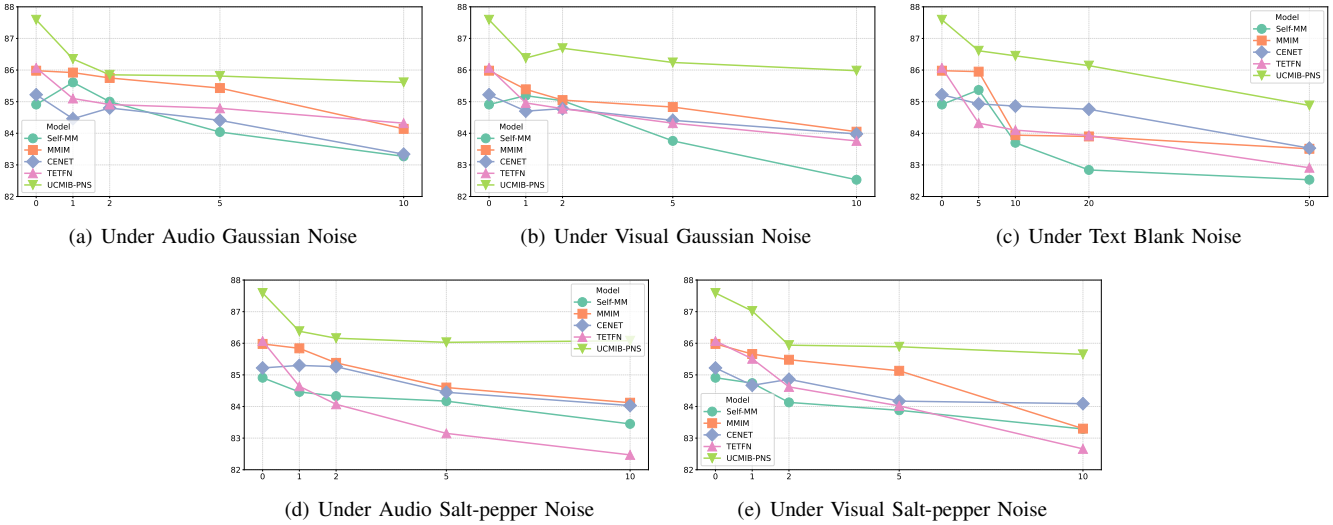
Fig. 8(a) illustrates the results on audio data corrupted by Gaussian noise. UCMIB-PNS consistently achieves the highest performance across varying noise levels. Moreover, its performance curve is relatively stable with a gentle slope, indicating strong robustness against noise in the audio modality. Fig. 8(b) presents the results on visual data under Gaussian noise conditions. Again, UCMIB-PNS outperforms all baseline models, maintaining a more stable trend with less performance degradation as noise increases. This demonstrates its superior noise resistance in the visual modality. Fig. 8(c) shows the results on blank text data with noise. Although all models experience some decline, UCMIB-PNS still leads in performance and exhibits the most stable curve, highlighting its robustness to noise in textual inputs. Fig. 8(d) shows the results for audio data with salt-and-pepper noise. UCMIB-PNS achieves the highest performance across all noise levels and maintains a nearly flat performance curve, indicating exceptional robustness against salt-and-pepper noise in the audio modality. In contrast, the baseline models exhibit greater fluctuations and performance degradation as noise intensity increases. Fig. 8(e) illustrates the performance on visual data with salt-and-pepper noise. Similar to the audio case, UCMIB-PNS consistently outperforms the baselines and demonstrates the most stable curve, suggesting strong resistance to visual perturbations caused by salt-and-pepper noise. Other models show more pronounced declines, confirming their sensitivity to such noise.

APPENDIX I
COMPLEXITY ANALYSIS

Table XX presents a comparison of model size, FLOPs, and performance (ACC-2 and F1) on the MOSI dataset. As shown in Algorithm 1, our method introduces no additional computational overhead during inference and adopts a text-centric dual-branch architecture instead of a pairwise design. Although UCMIB-PNS has slightly more parameters (86.139M) than the most compact model, Self-MM (85.806M), the difference is only 0.333M, which is negligible in practice. Its inference-time FLOPs remain the smallest, demonstrating comparable efficiency. UCMIB-PNS achieves the highest performance across all models, with significant improvements in both ACC-2 and F1 scores. Furthermore, it is worth noting that the FLOPs

Table XIX The sensitivity analysis of ϕ on noisy CMU-MOSI.

ϕ	Audio				Visual				Text
	Salt-pepper		Gaussian		Salt-pepper		Gaussian		Blank
	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=5	ϵ=10	ϵ=10
0.1	82.60/84.29	81.61/83.76	82.23/83.76	81.59/83.42	83.35/84.93	82.34/83.71	82.75/84.77	81.46/83.11	82.98/84.08
0.3	82.59/85.45	82.35/84.03	**83.74/85.81**	**83.70/85.61**	83.46/85.04	82.60/84.30	83.54/84.81	81.73/83.39	83.60/85.19
0.5	84.18/**86.08**	**84.00/86.08**	82.79/84.83	82.27/84.27	83.47/85.59	83.42/85.16	**84.15/86.24**	82.81/83.89	84.04/85.65
1.0	**84.27**/86.03	83.57/85.63	82.45/84.45	81.84/84.15	**83.82/85.89**	**83.59/85.65**	83.57/85.48	83.26/84.68	**84.34/86.45**
1.5	82.70/84.57	82.59/84.13	82.08/84.06	81.65/83.63	82.94/85.45	82.51/84.05	83.10/83.89	**83.74/85.98**	84.24/86.02



(a) Under Audio Gaussian Noise (b) Under Visual Gaussian Noise (c) Under Text Blank Noise

(d) Under Audio Salt-pepper Noise (e) Under Visual Salt-pepper Noise

Fig. 8: Visualization of Noise Robustness of UCMIB-PNS and Baseline Models on CMU-MOSI.

Table XX Comparison of Model Parameters and FLOPs. The marks are consistent with those used in Table I.

Model	Param (M)	FLOPs (G)	ACC-2	F1
MAG-BERT	86.870	4.312	82.37/84.43	82.50/84.61
Self-MM	**85.806**	4.274	82.54/84.77	82.68/84.91
MMIM	85.928	4.279	84.14/86.06	84.00/85.98
CENET	92.164	5.920	83.53/85.21	83.49/85.22
TETFN	86.612	4.298	84.05/86.10	83.83/86.07
UCMIB-PNS	86.139	**4.265**	**85.28/87.65**	**85.15/87.59**

remain unchanged even under noisy input conditions. The results confirm that UCMIB-PNS strikes a favorable balance between efficiency and robustness while delivering state-of-the-art performance.

APPENDIX J
COUNTERFACTUAL REASONING EVALUATION RESULTS

We employ two counterfactual reasoning approaches, counterfactual representations and counterfactual attention, to validate the effectiveness and robustness of our method.

A. Validation results based on counterfactual representations.

Many studies [59], [60], [88], [89] leverage counterfactual representations to reduce modality bias in multimodal learning, ensuring that models depend on causally relevant features rather than spurious correlations or the dominance of a single modality. Most methods construct counterfactual representations by intervening on one modality—typically by replacing its inputs or intermediate features with random tensors or scalars—while keeping the other modalities unchanged to simulate a "what-if" scenario. Following this common practice, in our experiment we create counterfactual scenarios by substituting the input of specific single modalities (Text, Audio, Vision) with random values. The experimental results are shown in Table XXI.

Table XXI presents the evaluation results on CMU-MOSI and CMU-MOSEI under counterfactual interventions. The model demonstrates noticeable robustness to perturbations in the vision and audio modalities, with only small deviations in prediction consistency even when these modalities are replaced with random noise (counterfactual scenario). This robustness can be attributed to the model's training strategy, which dynamically balances the sufficiency and necessity of each modality, preventing over-reliance on any single input source. When vision or audio signals become unreliable, the model adapts by increasing the relative weight of other modalities, thereby maintaining inference accuracy and ensuring outputs remain as consistent as possible with the original, non-intervened predictions. In contrast, interventions on the text modality lead to larger shifts, indicating that textual information still plays a dominant role in the final decision.

B. Validation results based on counterfactual attention

We follow prior work [61], [62], [90] by leveraging counterfactual attention to conduct counterfactual evaluation. Specifically, we intervene directly on attention, the key mediating

Table XXI Validation results based on counterfactual representations on CMU-MOSI and CMU-MOSEI datasets. $|\hat{y} - y'|$ measures the difference between the output after counterfactual intervention and the original model prediction. The table headers list the modalities undergoing counterfactual intervention.

Dataset	Text		Vision		Audio	
	F1	$\lvert\hat{y} - y'\rvert$	F1	$\lvert\hat{y} - y'\rvert$	F1	$\lvert\hat{y} - y'\rvert$
CMU-MOSI	76.14 / 76.64	0.652	83.30 / 85.19	0.315	83.64 / 85.23	0.307
CMU-MOSEI	71.78 / 80.63	0.386	84.74 / 86.24	0.158	85.10 / 86.70	0.141

Table XXII Validation results based on counterfactual attention on CMU-MOSI and CMU-MOSEI datasets. $|\hat{y} - y'|$ measures the prediction difference after intervention compared to the original model output. TA and TV denote counterfactual interventions on the Text-Audio and Text-Vision branches, respectively.

Method	CMU-MOSI				CMU-MOSEI			
	TA (F1)	TA ($\lvert\hat{y} - y'\rvert$)	TV (F1)	TV ($\lvert\hat{y} - y'\rvert$)	TA (F1)	TA ($\lvert\hat{y} - y'\rvert$)	TV (F1)	TV ($\lvert\hat{y} - y'\rvert$)
Random Attention	84.35 / 85.97	0.494	83.44 / 85.81	0.485	85.10 / 86.38	0.132	85.07 / 86.30	0.135
Uniform Attention	84.47 / 86.25	0.522	84.23 / 85.53	0.535	84.47 / 86.72	0.119	84.73 / 85.55	0.129
Reverse Attention	84.20 / 85.81	0.488	83.84 / 85.91	0.512	85.04 / 86.01	0.124	85.46 / 85.88	0.127
Shuffle Attention	83.96 / 86.20	0.509	83.38 / 85.44	0.524	85.39 / 86.61	0.117	84.57 / 86.77	0.130

variable in modality interactions, and design four types of counterfactual attention settings:

- Uniform attention: All unmasked positions are assigned an identical value equal to the average of the original attention weights, ensuring that each unmasked token shares the same weight.
- Reversed attention: Each unmasked position is assigned a value computed by subtracting its original attention score from the maximum attention value, effectively reversing the attention distribution across tokens.
- Shuffled attention: Attention weights are randomly permuted along the batch dimension.
- Random attention: Each unmasked position receives an attention score sampled uniformly from the range $\mathcal{U}(0, 2)$.

To ensure comprehensive evaluation, these interventions are applied separately to the text-audio branch and the text-vision branch. The experimental results for both branches are presented in Table XXII.

As shown in Table XXII, the F1 scores under all counterfactual attention settings remain close to the original performance on both CMU-MOSI and CMU-MOSEI datasets, and the prediction deviation $|\hat{y} - y'|$ is consistently low. This suggests that even when cross-attention patterns are randomized, reversed, or shuffled, the model's predictions are largely unaffected. The key reason for this robustness is the training strategy: by explicitly incorporating the notions of modality sufficiency and necessity, the model avoids over-reliance on any single modality. When counterfactual interventions disrupt the attention distribution of one modality branch, the model can dynamically reweight other modalities to maintain accurate predictions. This mechanism fundamentally enhances resilience against attention-level perturbations, demonstrating the model's strong capability to handle counterfactual attention shifts.

APPENDIX K
ADDITIONAL CASE ANALYSIS

To further enhance the interpretability of our model, we present a case analysis based on causal path analysis and cross-attention weight visualization on four representative test samples from the CMU-MOSEI dataset. This analysis helps to trace the influence of sufficient and necessary causal factors on the prediction results, providing a clearer understanding of the model's decision-making mechanism. Fig. 9 presents the results of the case analysis. The upper part illustrates the causal graph of our model when performing multimodal sentiment analysis, where T, A, and V represent the text, audio, and visual modalities, respectively, and TA and TV denote the branch representations after bimodal fusion. PNS indicates the results of reweighting the two branches using the PNS Estimator ($\mathrm{PNS} = \mathrm{PS} + \lambda\mathrm{PN}, \lambda = 1.2$ on CMU-MOSEI). Finally, TA and TV are multiplied by their corresponding weights to produce the final prediction. To further investigate how the causal effect operates within the model's decision-making mechanism, the lower part of Fig. 9 displays four cases from the same video segment, randomly selected from the CMU-MOSEI test set. For each case, we present the inputs from different modalities, the attention visualization of TA and TV, the weights estimated by the PNS Estimator, and the final prediction results. The attention visualization of TA and TV is obtained by averaging the cross-attention weights across all layers and all attention heads. For clearer visualization, PAD tokens in the text are removed, and the cross-attention weights are normalized using min-max scaling, followed by bicubic interpolation.

Specifically, in Case 1, the text modality contains sufficient semantic information that can easily resonate with either the audio or visual modality. However, we observe that the speaker in this case does not exhibit obvious sentimental expressions visually; therefore, the attention heatmap for the TV branch does not display prominent highlighted regions, indicating that high-level sentimental cues were not extracted. In contrast, the vocal tone exhibits a distinctly cold quality, expressing an undertone of disdain. Therefore, certain segments of the audio modality interact strongly with the text, and the highlighted regions reveal the positions of sentimental cues across the two modalities, demonstrating that the TA branch contains high-level sentimental information. Consequently, the PNS Estimator assigns a relatively high sufficiency weight to the TA branch, while the TV branch receives a lower sufficiency

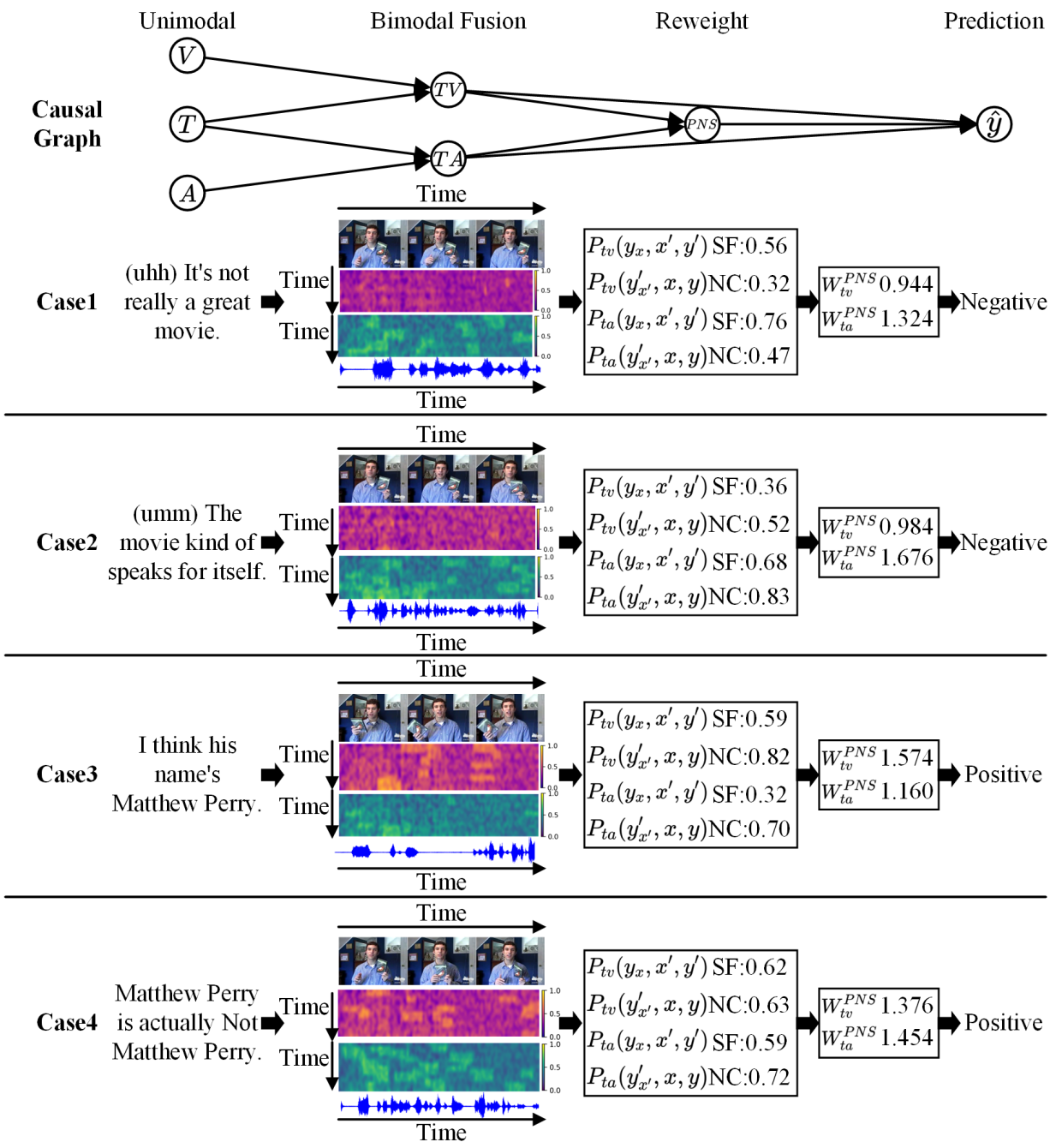


Fig. 9: The case analysis results based on causal path analysis and cross-attention weight visualization on four test samples from CMU-MOSEI.

weight. Moreover, since most sentimental cues are concentrated in TA, their necessity is also relatively high. Based on the above mechanism, the model ultimately makes a correct prediction that the sample expresses negative sentiment.

In Case 2, the sentimental semantics in the text modality are highly implicit, which limits its interaction with other modalities and makes it necessary to rely on sentimental cues from the audio or visual modalities for accurate sentiment prediction. Although the visual modality also lacks clear sentimental signals, the audio modality provides useful cues through tone and pauses. Consequently, similar to Case 1, the attention heatmap for the TV branch does not exhibit prominent highlighted regions, while noticeable highlights appear in specific audio segments due to the effective interaction between the implicit semantics of the audio and text modalities. As the textual semantic information diminishes, both TA and TV branches exhibit reduced sufficiency: the TV branch is strongly affected due to its reliance on textual cues, while the TA branch also declines because audio alone cannot fully capture sentiment. This situation causes the necessity of both TA and TV to increase substantially, reflecting their growing indispensability to each other. Based on this mechanism, the

model ultimately predicts the sample as expressing negative sentiment correctly.

In Case 3, the text modality contains almost no sentimental cues, presenting a relatively neutral expression. The speaker also exhibits minimal variation in vocal tone, resulting in an attention heatmap for the TA branch with no prominent highlights. However, the visual modality conveys a natural and pleasant demeanor, leading to strong resonance between text and video within the TV branch. Nevertheless, due to the scarcity of sentimental cues in the text, the fused representations of TA and TV contain fewer sentimental signals compared to Case 1 and Case 2. Since the visual modality provides sentimental cues, the TV branch exhibits higher sufficiency than TA. However, because neither branch alone demonstrates sufficient sufficiency, both need to increase their reliance on the other to make the final decision, indicating that both branches have high necessity. Based on this mechanism, the model ultimately predicts the sample as positive sentiment correctly.

Similar to Case 3, the text modality in Case 4 also lacks explicit sentimental cues. Fortunately, the speaker exhibits both a cheerful tone and a pleasant facial expression, which successfully interact with specific text tokens and are clearly reflected in the attention heatmaps. However, due to the absence of sentimental information in the primary text modality, the sufficiency of both branches remains low, requiring each branch to rely on the other for accurate prediction. In this situation, both branches actively seek complementary information from each other, reinforcing their interdependence. Consequently, both branches exhibit a certain degree of necessity. Based on this mechanism, the model ultimately predicts the sample as positive sentiment correctly.

The above case analysis on CMU-MOSEI offers an intuitive illustration of how our model dynamically integrates sufficiency and necessity in decision-making. Our findings reveal that sufficiency and necessity do not follow a simple inverse relationship; rather, their interaction is context-dependent and governed by the distribution of sentimental cues across modalities. When one modality lacks explicit sentiment signals, its sufficiency decreases; if the complementary modality provides discriminative sentimental information, its necessity increases because it becomes indispensable for the final prediction. Conversely, if neither modality offers strong cues, both exhibit low sufficiency and simultaneously higher necessity, requiring greater interdependence for accurate inference. In contrast, when sentimental cues are distributed across multiple modalities, sufficiency for each branch improves, and necessity correspondingly diminishes, as the prediction can be achieved with less reliance on other branches. These dynamics demonstrate that necessity is not triggered solely by the insufficiency of a branch but by the availability of compensatory cues in the complementary branch. By tracing causal paths, visualizing cross-attention patterns, and applying PNS-based reweighting, this analysis highlights how sufficiency and necessity probabilities jointly shape the prediction outcome, reinforcing the interpretability advantage of our causal framework.