Hyper-Opinion Vagueness Quantification for Robust Multimodal Learning

Disen Hu[1,2], **Xun Jiang**[2], **Xiaofeng Cao**[1],
Zheng Wang[1], **Jingkuan Song**[1], **Heng Tao Shen**[1], **Xing Xu**[1*]

[1]School of Computer Science and Technology, Tongji University, Shanghai, China
[2] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

Abstract

Robust Multimodal Learning (RML) aims to address the issues of unreliable predictions of multimodal models. Nevertheless, previous RML works often struggle to distinguish between different categories that rely on identical intra-modal cues, making ambiguous predictions. We defined this degree of "uncertain" in extracting discriminative features of a multimodal model as **vagueness**. Neglecting such vagueness, as previous RML works commonly do, will undermine the ability to extract unique semantics of each category in multimodal models, further resulting in worse robustness under disturbances that affect semantic representations. Additionally, this vagueness will lead the parameter updating processes towards unreliable fusion, thus diverting the learning processes of the multimodal model from learning unique features of each category. Based on the above insight, we propose a novel robust multimodal learning approach, termed ***H**yper-**O**pinion **V**agueness **Q**uantification (HOVQ)*. Specifically, we first introduce hyper-opinion to capture and quantify the vagueness of multimodal learning in discriminating representations of different categories. Moreover, to mitigate the interference in parameter updating of unreliable representations with high vagueness, we also design the Hyper-Opinion Gradient Modulation to guide the optimization processes. We evaluate our HOVQ on six datasets with different disturbances, including noise and adversarial attack, and demonstrate that our proposed method achieves state-of-the-art performance consistently.

Code — https://github.com/ConstantineWayne/HOVQ

Introduction

Multimodal models have achieved impressive success in all kinds of artificial intelligence applications (Lin and Hu 2023; Yang et al. 2023; Jiang et al. 2025; Wang et al. 2024; Jiang et al. 2024a; Hu et al. 2025a). However, in the presence of diverse disturbances, the performance of multimodal models with limited robustness can degrade significantly. Dedicated to this, Robust Multimodal Learning (Gao et al. 2024b; Zhang et al. 2023; Han et al. 2022), which aims to enhance the generalizability, robustness, and reliability of models, has gained increasing attention recently.

*Corresponding Author.

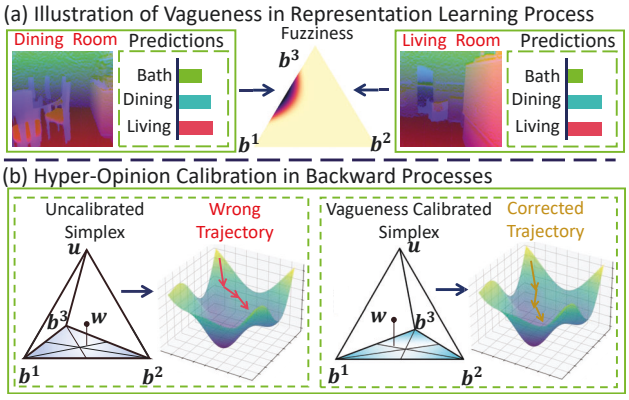

Figure 1: (a) Multimodal models heavily rely on identical intra-modal cues, further making unreliable and dissonant predictions. We defined this degree of "uncertain" to extract unique features of each category as **vagueness**. (b) Vagueness quantification would form a calibrated simplex, leading to a new optimization trajectory to a more robust solution. Conversely, biased belief mass without vagueness quantification will cause the model to be optimized on the wrong simplex. Noting that $b^1/b^2/b^3$ is the belief mass of the bath/-dining/living room, w is the opinion, and u is the uncertainty.

Accordingly, an essential motivation of RML lies in quantifying uncertainty inherent in data and incorporating it into the feature representation process to facilitate the learning of more distinguishable and reliable semantics. Specifically, the previous works (Gao et al. 2024b,a; Zhang et al. 2023; Han et al. 2022; Huang et al. 2025) are dedicated to acquiring uncertainty from the feature distributions and collecting confidence for the corresponding category, using techniques such as Gaussian projection and Evidential Deep Learning. Although their method showed some effectiveness, they **overlooked the fact** that due to insufficient semantic decoupling, the representations from different categories may heavily rely on identical semantic features, resulting in **vagueness** in differentiating representative information, producing decision-making with fuzzy probabilities or beliefs. When confronted with disturbances that obscure semantic information, multimodal models may fail to extract effective features for prediction, largely due to their low ro-

bustness resulting from the neglect of vagueness.

To further reveal such vagueness, we conducted empirical experiments using ECML (Xu et al. 2024) on NYU Depth v2. As shown in Fig. 1 (a), there exists a significant conflict in distinguishing "Dining Room" and "Living Room". This confusion arises from an over-coupling between these two categories, as both heavily rely on shared semantic cues such as "chair" and "table". Consequently, although the model can successfully exclude irrelevant categories like "Bathroom", it fails to confidently decide between "Dining Room" and "Living Room".

Additionally, the previous works consistently neglected the above vagueness, leading to a biased probability space and forming an unreliable simplex. This unreliable simplex with uncalibrated belief will disturb the parameter updating trajectory in backward processes. As the model struggles to disentangle class-discriminative representations and instead oscillates between overlapping semantic cues, the optimization without correction will further weaken the model's ability to extract unique category-wise features. As shown in Fig. 1(b), the updating processes without the calibration of vagueness will misguide the whole optimization trajectory, leading to unreliable learning and ultimately degrading both the generalization ability and robustness of the model.

To address the issue of vagueness brought by the over-coupling in multimodal models, *aligning with the motivation of RML, the vagueness should be introduced as a calibration during representation learning processes. Thus, the key lies in capturing and quantifying the vagueness.* In detail, we introduced hyper-opinion in Evidential Deep Learning, which is defined over an extended hyper-domain that includes both singleton categories and composite category subsets. *That is to say, hyper-opinion explicitly captures and quantifies vagueness by assigning belief mass to composite sets, groups of plausible classes that reflect semantic ambiguity, within a balanced simplex*. To this end, we proposed *Hyper-Opinion Vagueness Quantification (HOVQ)*. Specifically, we utilized the Grouped Dirichlet Distribution to model the hyper-opinion and quantify the vagueness in representations of multimodal models. Moreover, we designed the Hyper-Opinion Gradient Modulation to dynamically modulate the parameter updating trajectory, getting rid of the impact of unreliable learning processes. We evaluated our HOVQ on six datasets with different kinds of disturbances and achieved state-of-the-art performance, showing significant robustness and reliability.

Overall, our contributions are summarized in threefold:

- To diminish the impact of the vagueness that stems from the over-coupling issue in multimodal learning, we leveraged the hyper-opinion to capture and quantify it, enhancing the robustness of the models.
- We designed Hyper-Opinion Gradient Modulation to avoid the inference of unreliable representations without vagueness calibration in optimization processes.
- We evaluate our model on six widely used multimodal datasets with natural noises and adversarial attacks, gaining state-of-the-art and robust performances.

Related Work

Evidential Deep Learning. Evidential Deep Learning (EDL) is a probability framework that extends conventional classification or regression (Bao, Yu, and Kong 2021; Zhao et al. 2020; Fu et al. 2023; Amini et al. 2020; Pandey and Yu 2023; Wang et al. 2025) by modeling prediction uncertainty through evidence theory. This framework allows models to simultaneously output both beliefs in specific classes and the uncertainty about these beliefs. Evidential Deep Learning proved to be effective in Out-of-distribution (Jiang et al. 2024b; Xu et al. 2025; Huang et al. 2025; Liu, Chen, and Yue 2025) and anomaly detection (Chen et al. 2022; Hu et al. 2021; Bao, Yu, and Kong 2021). Moreover, to integrate the multi-view or multimodal in EDL, the Dempster-Shafer Evidence Theory (DST) (Dempster 1968, 2008) is employed, which directly models the uncertainty in subjective probabilities. In detail, Dempster's rule is able to fuse the shared parts of the heterogeneous sources and ignores conflicting beliefs. However, most existing Evidential Deep Learning can only quantify uncertainty through evidence that supports a single category, making it incapable of capturing the vagueness in representation learning processes. In this work, we introduced the hyper-opinion (Qu et al. 2024; Li et al. 2024) to avoid such insufficient uncertainty modeling, forming reliable multimodal learning.

Robust Multimodal Learning. Multimodal learning exhibited outstanding performances by integrating multiple information sources (Hu et al. 2025b; Liu et al. 2025; Guo et al. 2024; Baltrušaitis, Ahuja, and Morency 2018; Chaptoukaev et al. 2024). However, in real-world scenarios, multimodal data may suffer from degeneration (*e.g.*, image noises, modality missing, adversarial attacks *etc.*). Many researchers aim to enhance multimodal models' robustness to confront these disturbances: capturing and quantifying aleatoric and epistemic uncertainty through projecting them into the Gaussian space (Gao et al. 2024b,a; Wei et al. 2024), dynamic fusion to constrain the generalization boundary (Zhang et al. 2023; Han et al. 2022), and modulating training processes (Yang et al. 2024). Through utilizing the uncertainty estimation, the previous works have achieved promising results on robust multimodal learning by integrating the uncertainty produced by the model into the final decision-making. Regrettably, they neglected that the biased uncertainty quantification can not only harm the confident prediction during the feed-forward process, but also can influence the models' optimization processes in backward processes. Thus, we calibrate the uncertainty to adjust the parameter updating.

Method

Preliminary of Hyper-Opinion

Traditional Evidential Deep Learning assigns beliefs to singleton categories based on multinomial opinion of Subjective Logic within domain $\mathbb{X}$ with cardinality of K using Dirichlet distribution. However, in domain $\mathbb{X}$, a limited portion of the hyper-domain, Evidential Deep Learning can only offer the sharp belief masses estimation, which will reduce the model's representative abilities due to the vague-

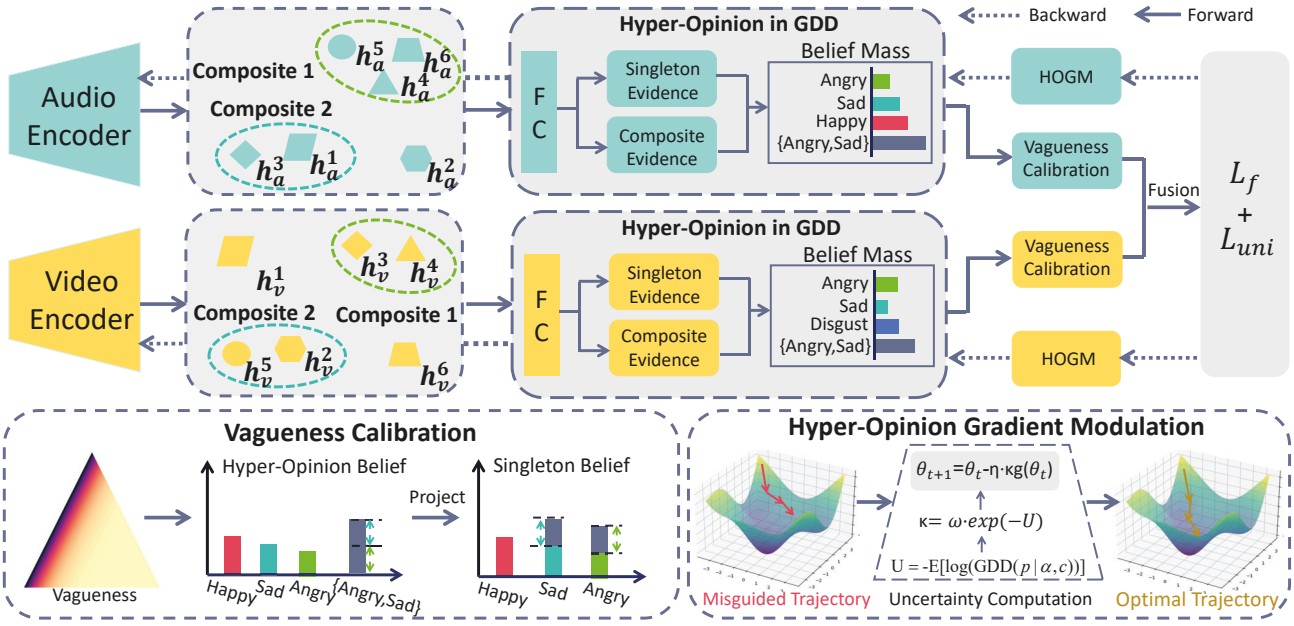

Figure 2: HOVQ aims to quantify the vagueness in learning processes and use it to calibrate the representation learning. **For quantification**, the composites are constructed based on category features centroids, and Grouped Dirichlet Distribution is employed to capture the vagueness based on the category composites. **For calibration,** the vagueness is projected onto singletons to calibrate the biased belief distributions, while Hyper-Opinion Gradient Modulation is designed to guide the learning process.

ness brought by over-coupling issues. In order to avoid the biased uncertainty modeling, we model the evidence in the *hyper-domain* $\mathcal{R}(\mathbb{X})$ to form the hyper-opinion:

$$\mathcal{R}(\mathbb{X}) = \mathcal{P}(\mathbb{X})/\{\{\mathbb{X}\},\{\emptyset\}\}, \tag{1}$$

where $\mathcal{P}(\mathbb{X})$ denotes the power set. We also defined the set of composite sets $\mathcal{C}(\mathbb{X}) = \mathcal{R}(\mathbb{X})/\{\{1\},\ldots,\{K\}\}$.

In a *hyper-opinion* $\boldsymbol{w} = (\boldsymbol{b}, u)$, a belief mass $b_{\mathcal{S}}$ is allocated to each subset $\mathcal{S} \in \mathcal{R}(\mathbb{X})$ (singleton class or composite set), along with an uncertainty mass u. These values are all non-negative and constrained to sum to one, *i.e.*,

$$\sum_{\mathcal{S}\in\mathcal{R}(\mathbb{X})} b_{\mathcal{S}} + u = 1. \tag{2}$$

Particularly, the belief $b_{\mathcal{S}}$ and the uncertainty mass u are formulated as:

$$b_{\mathcal{S}} = \frac{e_{\mathcal{S}}}{T}, \quad u = \frac{K+\sigma}{T}, \tag{3}$$

where σ is the number of the composite sets, while $e_{\mathcal{S}}$ is the non-negative evidences derived for $\mathcal{S}$ and $T = \sum_{\mathcal{S}\in\mathcal{R}(\mathbb{X})} e_{\mathcal{S}} + K + \sigma$. And the belief masses reside on the probability simplex, where each vertex corresponds to a distinct class, which is similar to Fig. 1, for maintaining semantic separability between categories.

Vagueness Quantification Through Hyper-Opinion

The pipeline of our HOVQ is shown in Fig. 2, since our motivation is to capture and quantify the vagueness that stems from the over-coupling of the features, we need to characterize the category combinations that reflect the vagueness.

Composites Construction. Supposing there are two modalities: A and V, we denote the features after the encoders are $\mathbf{F_a}, \mathbf{F_v} \in \mathbb{R}^d$ respectively (d is the hidden dimension of the modality features). The subsequent functions are formulated from the perspective of the video modality, while the audio modality follows an analogous formulation.

Firstly, we clustered the features based on the centroids $\mathbf{h_v^k}$ of each category k:

$$\mathbf{h_v^k} = \frac{1}{N_k} \sum_{j:y_j=k} \mathbf{F_v^j}. \tag{4}$$

And then we calculate the similarities between different categorical centers:

$$\mathrm{Sim}_v = \frac{\mathbf{h_v^{k_1}} \cdot \mathbf{h_v^{k_2}}}{\|\mathbf{h_v^{k_1}}\| \cdot \|\mathbf{h_v^{k_2}}\|}, \quad k_1, k_2 \in [1, K]. \tag{5}$$

After that, if the similarities are higher than the threshold γ, we consider that these categories can construct a composite. This process allows the model to reason over semantically correlated label subsets and supports the learning of vague class concepts.

Vagueness Quantification. Traditional Dirichlet Distribution used in Evidential Deep Learning can only form a sharp belief to capture the aleatoric uncertainty, failing to fully express the hyper-opinion. As shown in Fig. 3, the hyper-opinion can explicitly capture the vagueness caused by over-coupling through category composites. We now introduce the Grouped Dirichlet Distribution (GDD)–a hyper-Dirichlet Distribution–to characterize the vagueness in multimodal models through the above-constructed composites

and corresponding singleton categories. In detail, the prerequisites for GDD is that: composite sets in $\mathcal{C}(\mathbb{X})$ form a partition of the set of singleton classes $\mathbb{X}$, *i.e.*, the composite set $\mathcal{S}^c = \{\mathcal{S}_1^c, \ldots, \mathcal{S}_\sigma^c\}$, where $\bigcup_{i=1}^{\sigma} \mathcal{S}^c = \mathbb{X}$ and $\mathcal{S}_j^c \bigcap S_i^c = \emptyset, \forall i, j \in \{1, \ldots, \sigma\}$ and $i \neq j$. For simplicity, we denote the composite sets S^c of audio and video as $\mathcal{A}$ and $\mathcal{V}$, respectively.

To acquire the GDD for each modality, i.e., $GDD_v(\boldsymbol{\alpha^v}, \boldsymbol{c^v})$, we first calculate the class-specific concentration parameters $\boldsymbol{\alpha^v}$ and composite-specific concentration parameters $\boldsymbol{c^v}$ through transforming the unimodal evidence vector $\boldsymbol{e^v} = [e_1^v, \ldots, e_K^v, e_{\mathcal{V}_1}^v, \ldots, e_{\mathcal{V}_\sigma}^v]$, where $\mathcal{V}_\sigma$ is the σ^{th} composite. And the evidence vector is acquired by the following linear transformation:

$$\mathbf{e}^v = \log(1 + \exp(\mathbf{W}\mathbf{F}_v + \mathbf{b})), \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{(K+\sigma)\times d}$ is learnable weight matrix, and $\mathbf{b} \in \mathbb{R}^{K+\sigma}$ is the learnable bias. And the concentration parameters of video modality are defined as:

$$\boldsymbol{\alpha^v} = [\alpha_1^v, \ldots, \alpha_K^v] = [e_1^v + 1, \ldots, e_K^v + 1], \tag{7}$$

$$\boldsymbol{c^v} = [c_{\mathcal{V}_1}^v, \ldots, c_{\mathcal{V}_\sigma}^v] = [e_{\mathcal{V}_1}^v, \ldots, e_{\mathcal{V}_\sigma}^v]. \tag{8}$$

Moreover, according to Qu *et al.* , (Qu et al. 2024), the belief mass assigned to a composite set $\mathcal{C}(\mathbb{X})$ is the vagueness V of each modality:

$$V^v = \sum_{j=1}^{\sigma} b_{\mathcal{V}_j}^v, \quad b_{\mathcal{V}_j}^v = \frac{e_{\mathcal{V}_j}^v}{T}. \tag{9}$$

Based on the above equations, the probability density function (PDF) of video GDD is:

$$GDD_v(\mathbf{p}|\boldsymbol{\alpha^v}, \boldsymbol{c^v}) = Z^{-1}(\prod_{k=1}^{K} p_k^{\alpha_k^v - 1})(\prod_{j=1}^{\sigma}(\sum_{l\in\mathcal{V}_j} p_l)^{c_{\mathcal{V}_j}^v}). \tag{10}$$

Moreover, the Z serves as the normalization constant of GDD, and is defined in the above equation:

$$Z = [\prod_{j=1}^{\sigma} B(\{\alpha_l^v\}_{l\in\mathcal{V}j})]B(\{\beta_j\}_{j=1}^{\sigma}), \beta = \sum_{l\in\mathcal{V}_j} \alpha_l + c_{\mathcal{V}_j}, \tag{11}$$

where $B(\cdot)$ is the beta function.

Additionally, the loss function of constructing GDD has two parts:

$$\mathcal{L}_{kl}^v = KL(GDD_v(\mathbf{p}|\bar{\boldsymbol{\alpha}}^{\boldsymbol{v}}, \bar{\boldsymbol{c}}^{\boldsymbol{v}})||GDD(\mathbf{p}|\mathbf{1}^K, \mathbf{0}^\sigma)), \tag{12}$$

$$\mathcal{L}_{PCE}^v = \mathbb{E}_{\mathbf{p}\sim \mathrm{GDD}_v(\mathbf{p}|\boldsymbol{\alpha}^v, \boldsymbol{c}^v)}\left[-\log\left(\tilde{\mathbf{y}}^\top \mathbf{p}\right)\right], \tag{13}$$

in which KL is the KL-divergence which regulates the distribution to make the evidence output more flat, $\tilde{y}$ is the one-hot label (*e.g.*, $\tilde{y}^{(i)} = [1, 1, 0]$ denotes composite categories of {1,2}) and $\bar{\boldsymbol{\alpha}}^{\boldsymbol{v}} = \tilde{y} + (1 - \tilde{y}) \odot \boldsymbol{\alpha^v}$, $\bar{\boldsymbol{c}}^{\boldsymbol{v}} = (1 - \tilde{y}) \odot \boldsymbol{c^v}$. And $\mathbf{1}^K$ denotes a K-dimensional vector of ones, and $\mathbf{0}^\sigma$ denotes a σ-dimensional vector of zeros. Finally, the total loss for unimodal GDD is:

$$\mathcal{L}_{GDD}^v = \mathcal{L}_{PCE}^v - \lambda \mathcal{L}_{KL}^v, \tag{14}$$

and λ is the hyperparameter to be decided.

Vagueness Calibration. In the previous paragraph, we successfully modeled the hyper-opinion using GDD, and quantified the vagueness of each modality through Eq. 9. We need to empower our model with the ability to consider such vagueness to calibrate the traditional uncertainty. Since the hyper-opinion allows multiple singletons in $\mathbb{X}$ to be considered to be true at simultaneously, the vague belief mass can be projected to the corresponding singletons, integrating the vagueness into traditional uncertainty quantification:

$$V_k^v = \sum_{l\in\mathcal{V}_j} \delta_{l=k} \cdot a(k \mid l) \cdot b_l,$$
$$a(k \mid l) = \frac{1}{|\mathcal{V}_j|}, k \in [1, K], \quad j \in [1, \sigma], \tag{15}$$

where V_k^v is the projected video vagueness onto the singleton category k, while $\delta_{l=k}$ is the Kronecker delta, which equals 1 if $l = k$ and 0 otherwise.

Consequently, the more comprehensive evidence capturing which incorporates the vagueness calibration is:

$$b_k^v \leftarrow b_k^v + V_k^v, \quad k \in [1, K]. \tag{16}$$

Through the above equation, the vague beliefs can successfully integrate into the singletons' belief masses of video b_k^v to enhance the model's vague cognitive abilities, gaining a more reliable unimodal opinion $w_v = (\boldsymbol{b^v}, u^v)$ with an unbiased belief distribution over the simplex, estimated through the Dirichlet Distribution derived from the video modality.

Fusion Method. To fuse the multiple Dirichlet Distribution of different modalities, following previous works in EDL (Han et al. 2022; Xu et al. 2024), we adapted the Dempster's combination rule to gain the combination mass of unimodal opinions (*i.e.*, $w_v = (\boldsymbol{b^v}, u^v)$ and $w_a = (\boldsymbol{b^a}, u^a)$), forming a joint opinion $\boldsymbol{w}_f = (\boldsymbol{b^f}, u^f)$ along with the joint Dirichlet Distribution $D(\boldsymbol{p^f}|\boldsymbol{\alpha^f})$:

$$b_k^f = \frac{1}{1-R}(b_k^a b_k^v + b_k^a u^v + b_k^v u^a), u^f = \frac{1}{1-R} u^a u^v, \tag{17}$$

where $R = \sum_{i\neq j} b_i^a b_j^v$ quantifies the degree of disagreement between the two mass distributions, and the scalar factor $\frac{1}{1-R}$ is used for normalization.

Hyper-Opinion Gradient Modulation

We argue that without properly vagueness quantification, the model generates biased belief masses over the probability simplex, leading to unreliable supervision signals. This misguides the optimization process and weakens the ability to learn discriminative multimodal representations. As a result, we consider modulating the model's learning pace based on the uncertainty of each feed-forward iteration in the **backward processes**. Specifically, we calculate the model's uncertainty in each iteration for each modality as:

$$U_v = -\mathbb{E}[log(GDD_v(\mathbf{p}|\boldsymbol{\alpha^v}, \boldsymbol{c^v}))]. \tag{18}$$

It is the same for U_a. Instead of directly using uncertainty in Eq. 3, the above is the entropy of GDD, which represents the uncertainty of both singletons and composites, pro-

Dataset	Method	Clean	Gaussian Noise		Salt-pepper	
		$\epsilon = 0.0$	$\epsilon = 5.0$	$\epsilon = 10.0$	$\epsilon = 5.0$	$\epsilon = 10.0$
MVSA	QMF (ICML'23)	78.07	73.85	61.28	73.90	60.41
	EAU (CVPR'24)	79.15	73.34	61.78	73.69	60.46
	MMPareto (ICML'24)	64.16	52.05	45.15	60.31	54.33
	ECML (AAAI'24)	76.83	71.28	61.03	72.13	61.04
	CRMT (ICLR'24)	65.33	53.78	45.33	55.62	44.54
	NLC (AAAI'25)	73.79	65.39	58.98	66.64	57.28
	HOVQ (Ours)	**81.31**	**74.71**	**63.34**	**74.25**	**61.81**
NYU Depth V2	QMF (ICML'23)	70.09	61.62	55.60	58.50	45.69
	EAU (CVPR'24)	72.05	62.54	56.23	58.44	46.21
	MMPareto (ICML'24)	71.67	60.55	53.32	55.81	44.18
	ECML (AAAI'24)	71.72	62.08	54.58	57.57	44.93
	CRMT (ICLR'24)	66.80	55.93	45.43	54.66	43.12
	NLC (AAAI'25)	67.33	54.90	45.04	56.02	44.66
	HOVQ (Ours)	**74.17**	**64.67**	**57.43**	**61.59**	**47.16**
SUN RGB-D	QMF (ICML'23)	61.98	53.40	48.58	52.49	40.53
	EAU (CVPR'24)	55.68	49.39	44.23	50.38	38.38
	MMPareto (ICML'24)	57.89	48.12	43.77	49.90	38.91
	ECML (AAAI'24)	59.82	52.46	46.71	51.92	39.29
	CRMT (ICLR'24)	50.32	41.37	35.33	42.18	34.70
	NLC (AAAI'25)	52.75	43.57	38.49	45.07	37.25
	HOVQ (Ours)	**62.24**	**55.20**	**49.78**	**53.63**	**41.60**
CREMA-D	QMF (ICML'23)	70.30	61.16	53.00	63.17	57.07
	EAU (CVPR'24)	68.30	60.90	54.58	61.49	55.32
	MMPareto (ICML'24)	75.13	54.16	39.65	67.87	59.04
	ECML (AAAI'24)	66.85	57.88	50.03	65.86	56.55
	CRMT (ICLR'24)	69.04	60.26	53.58	62.11	55.76
	NLC (AAAI'25)	64.79	55.98	47.31	58.47	52.77
	HOVQ (Ours)	**75.19**	**67.82**	**63.94**	**71.48**	**67.98**

Table 1: Comparison with SOTA methods on four datasets when 50 % of modalities are corrupted with two kinds of noises (*i.e.*, Gaussian and Salt-pepper). Note that we reported the **average accuracy** over five different random seeds.

viding a more fine-grained method to quantify the uncertainty for hyper-opinion in Evidential Deep Learning. Additionally, to adjust the learning through the backward processes, we leveraged the gradient modulation technique–Hyper-Opinion Gradient Modulation (**HOGM**):

$$\kappa_a = \omega_a \cdot exp(-U_a), \kappa_v = \omega_v \cdot exp(-U_v), \tag{19}$$

where ω_a, ω_v are the temperature factors. The higher the uncertainty is, the smaller the gradient weight κ is, meaning that the model should not slow its pace in this uncertain iteration, and vice versa.

The second part of HOGM to modulate gradients is:

$$\boldsymbol{\theta}^{\boldsymbol{u}}_{\boldsymbol{t+1}} = \boldsymbol{\theta}^{\boldsymbol{u}}_{\boldsymbol{t}} - \eta \cdot \kappa_u g(\boldsymbol{\theta}^{\boldsymbol{u}}_{\boldsymbol{t}}), \quad u \in \{A, V\}. \tag{20}$$

The above equation is the final parameter updating process for modality u, in which $\boldsymbol{\theta^u}$ is the unimodal model parameters, η is the learning rate, and $g(\cdot)$ stands for the gradients of parameters $\boldsymbol{\theta^u}$.

Loss Function

After gaining the joint belief mass distribution through Eq. 17, the classification loss for this distribution is:

$$\begin{aligned}\mathcal{L}_f = &\mathbb{E}_{\boldsymbol{p^f} \sim D(\boldsymbol{p}^f | \boldsymbol{\alpha}^f)} \left[\log p(\tilde{\boldsymbol{y}} | \boldsymbol{p}^f)\right] \\ &- \lambda_j \, \mathrm{KL}\left(D(\boldsymbol{p^f} | \tilde{\boldsymbol{\alpha}}^{\boldsymbol{f}}) \,\|\, D(\boldsymbol{p}^f | \mathbf{1}^K)\right),\end{aligned} \tag{21}$$

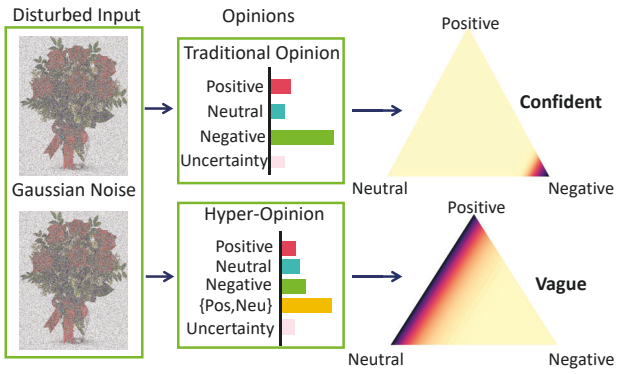

Figure 3: Unlike the traditional opinion, the hyper-opinion can depict the vagueness of the model in distinguishing different categories due to insufficient semantics decoupling.

where λ_j is the trade-off parameter. We also have two losses for unimodal GDD distribution following Eq. 14:

$$\mathcal{L}_{uni} = L^a_{GDD} + L^v_{GDD}. \tag{22}$$

Thus, the overall loss of our method is:

$$\mathcal{L}_{total} = \mathcal{L}_{uni} + \mathcal{L}_f. \tag{23}$$

Experiments

Experimental Setups

Datasets. We conducted experiments on six datasets. ∘ MVSA-Single (Niu et al. 2016) includes multiple image-text pairs with manual sentimental annotations. ∘ NYU Depth V2 (Silberman et al. 2012) comprises rgb and depth images for indoor classification. ∘ SUN RGB-D (Song, Lichtenberg, and Xiao 2015) is another indoor scene recognition dataset, containing 19 classes. ∘ CREMA-D (Cao et al. 2014) contains acoustic and visual modalities and is used for sentiment analysis. ∘ Kinetics-Sounds (Kay et al. 2017) encompasses 19,000 video clips categorized in 31 categories. ∘ UCF101 (Soomro, Zamir, and Shah 2012) aims to predict human motions using RGB images and optical flows.

Implementation Details. We followed the previous works to implement noises (varying noise rate ϵ from 5.0 to 10.0) (Zhang et al. 2023) and attacks (ℓ_2 PGD (Madry et al. 2017) and FGM (Goodfellow, Shlens, and Szegedy 2014) attack with attack size $\epsilon = 0.5$) (Yang et al. 2024). We set a batch size of 16 on each dataset and train our method within 150 epochs with the learning rate of $1e^{-2}$. We used the SGD optimizer for CREMA-D, Kinetics-Sounds, and UCF101 datasets, and Adam for the rest. We set the threshold γ to 0.5 for every datasets and the setting of component number can be found in our code.

Comparison With SOTA Methods

In order to test our method's robustness under different kinds of data disturbances, we compared our method with other methods that aim to enhance the multimodal model's robustness (TMC(Han et al. 2022), QMF(Zhang et al. 2023), EAU(Gao et al. 2024a), MMPareto(Wei and Hu 2024), CRMT (Wei et al. 2024),). (Among them, except for on KS and UCF101, the experiments of MMPareto and CRMT are reproduced according to their official code.) Some of the methods (ECML(Xu et al. 2024), NLC(Xu et al. 2025)) for robust multi-view classification are also adapted (methods are reproduced according to their open-source code). The results are listed in Table. 1. It is obvious to notice that our proposed method can reach state-of-the-art performances across all four multimodal datasets on both clean and noisy conditions. To be specific, our method can surpass its counterparts with around 3-4 % improvements. Compared to QMF and EAU, which are typical methods for robust multimodal learning, our method can gain almost 2-3 % enhancement in two scenarios.

Additional Evaluation

Unlike the natural disturbances, ℓ_2 PGD and FGM attacks are two adversarial attacks that may lead to catastrophic collapse of a deep learning model, and are also widely used to test the model's robustness. Following previous work (Wei et al. 2024), we adapted the above two attacks on Kinetics Sounds and UCF101 datasets to demonstrate our model's robustness against the adversarial attack; the results are shown in the Table. 2. It is clear that even with these two types of adversarial attacks, our model exhibits outstanding robustness. Our method can surpass CRMT–the lasted method for robust multimodal learning against adversarial attack–with an improvement of around 2% with and without attack.

Dataset	Method	Clean	ℓ_2 PGD	FGM
UCF101	QMF (ICML'23)	0.754	0.536	0.550
	EAU (CVPR'24)	0.589	0.365	0.321
	MMPareto (ICML'24)	0.744	0.501	0.465
	CRMT (ICLR'24)	0.759	0.614	0.602
	HOVQ (Ours)	**0.777**	**0.625**	**0.612**
KS	QMF (ICML'23)	0.768	0.591	0.560
	EAU (CVPR'24)	0.704	0.423	0.408
	MMPareto (ICML'24)	0.701	0.444	0.416
	CRMT (ICLR'24)	0.762	0.608	0.602
	HOVQ (Ours)	**0.780**	**0.649**	**0.616**

Table 2: Comparison with state-of-the-art robust multimodal learning methods under ℓ_2 PGD and FGM attack. We report adversarial accuracy results. KS stands for Kinetics-Sounds.

HOGM	$\mathcal{L}_{uni}$	$\mathcal{L}_f$	$\epsilon = 0$ Acc	$\epsilon = 5.0$ Acc	$\epsilon = 10.0$ Acc
✓	✗	✓	77.45	71.86	61.65
✓	✓	✗	**79.57**	**74.37**	**63.77**
✗	✓	✓	77.84	70.05	53.95
✗	✗	✓	68.78	52.79	49.79
✗	✓	✗	75.42	71.09	51.25

Table 3: Ablation study on MVSA-Single with different Gaussian noises.

Further Analysis

Ablation Study. We compared different combinations of components in our proposed method (HOGM, $\mathcal{L}_{uni}$, and $\mathcal{L}_f$). Here we exhibit the experiment results on MVSA-Single with different Gaussian noises. It is clear to see, when any of the components is ablated, the model's performance drops significantly (around 5 %) under clean and noisy conditions. Specifically, with the absence of HOGM, our model suffers from a dramatic drop of 7 %, meaning its important role in enhancing the model's robustness with HOGM modulation, which helps the model to update parameters according to the uncertainty after the one-step training. Moreover,

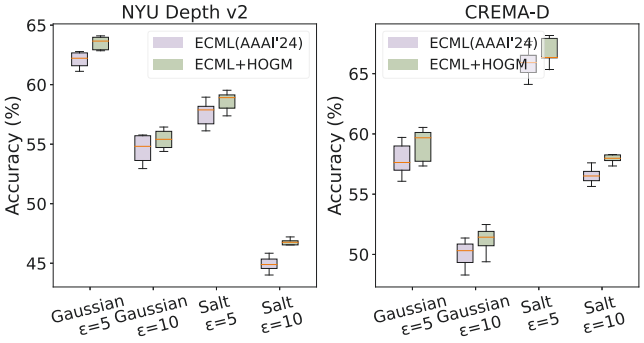

Figure 4: Performances of ECML combined with our HOGM module on CREMA-D and NYU Depth v2 datasets with different noises and intensities.

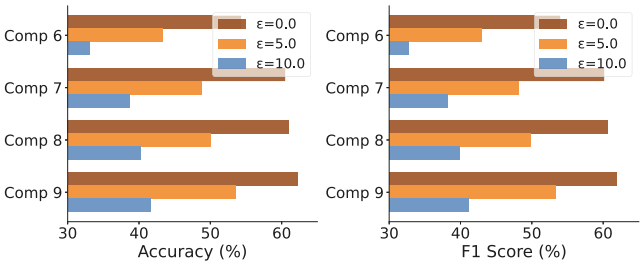

Figure 5: The ablation study on the number of composites on SUN RGB-D with different degrees of noise.

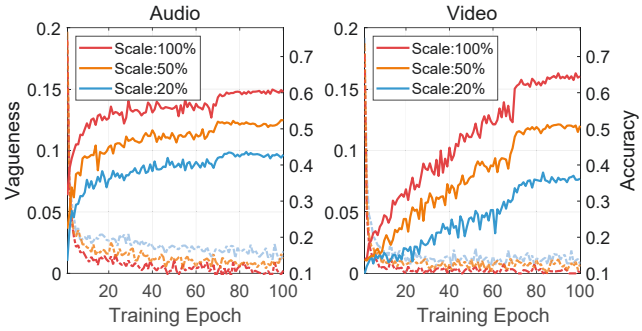

Figure 6: The experiments on different scales of training set. We represented the unimodal vagueness and performances on Kinetics-Sounds. The solid and dashed lines represent the uni-modal accuracy and the vagueness, respectively.

we can observe that $\mathcal{L}_{uni}$ also facilitates the model to learn from vague composite, when it is removed, the performance of HOVQ degenerates rapidly.

Ablation Study on Composites Number. A number of composites play an important role in facilitating the learning from vague evidence; thus, in order to verify their impacts, we conducted experiments with different numbers of composites on SUN RGB-D under clean and **Salt-pepper** noise conditions, shown in the following figure. It is clear to see that with the increasing number of composites, the model's performance improved as well. This indicates that a higher number of composites means that the model has richer sources of evidence, making the final predicted belief more robust and less uncertain.

Effectiveness and Generalizability of HOGM. In order to testify the importance of introducing uncertainty into backward processes to guide the models' learning, we add our HOGM module to the ECML and conducted experiments on two datasets with different noises and intensities. The results are shown in Fig. 4. It is obvious to see that with the help of HOGM, the performances of ECML on both datasets with various noises can be enhanced by around 1%. It indicates that the modulation on the model's learning trajectory can lead the model to learn the optimal parameter updating directions, resulting in better performance.

Discussion on Vagueness During Training. Based on our motivation, we visualize vagueness fluctuation and unimodal accuracy on Kinetics-Sounds, showing the benefit of

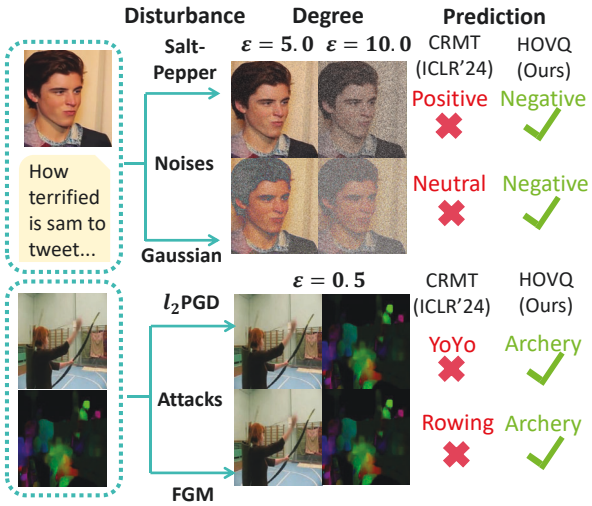

Figure 7: Visualizations of test cases selected from the MVSA-single and UCF101.

leveraging vagueness (Fig. 6). The dataset scales range from 100% to 20%. With a fixed test set, reducing training size naturally introduces vagueness. During training, our model learns from composite-based vague evidence and gradually improves confidence in singleton predictions, boosting performance. As vagueness decreases, acoustic and visual accuracies increase accordingly. Compared to the full training set, subsets maintain higher vagueness throughout training, leading to poorer results. Since vagueness is inversely correlated with performance, collecting sufficient evidence reduces it and enhances both unimodal and multimodal results.

Quantitative Analysis. Additionally, we also illustrate several representative test cases from the MVSA-Single dataset and the UCF101 dataset with noise and adversarial attacks. As shown in Fig. 7, on the MVSA-single dataset with different noises, the counterpart outputs false predictions, while our method can perfectly fulfill the classification task. Moreover, on the UCF101 dataset, when the two modalities are attacked by two different adversarial attacks, CRMT is unable to make the right classifications. Meanwhile, HOVQ can consistently output the correct predictions.

Conclusion

In this work, we proposed a novel framework–HOVQ to enhance the multimodal models' robustness in the face of various kinds of disturbances. In detail, our method can learn from the vague evidence through the composite categories, and can also modulate the model's learning pace based on the uncertainty in each iteration in the backward processes. With the modulation in both forward and backward processes, our model can show outstanding robustness and generalizability. We evaluate our method on six widely used multimodal datasets with various disturbances, and our method can reach state-of-the-art performances.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants, China (No.62476201, 62222203 and 62306065), New Cornerstone Science Foundation through the XPLORER PRIZE, and Meituan.

References

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in neural information processing systems*, 33: 14927–14937.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Bao, W.; Yu, Q.; and Kong, Y. 2021. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13349–13358.

Cao, H.; Cooper, D. G.; Keutmann, M. K.; Gur, R. C.; Nenkova, A.; and Verma, R. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4): 377–390.

Chaptoukaev, H.; Marcianó, V.; Galati, F.; and Zuluaga, M. A. 2024. Hypermm: Robust multimodal learning with varying-sized inputs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 170–183.

Chen, L.; Lou, Y.; He, J.; Bai, T.; and Deng, M. 2022. Evidential neighborhood contrastive learning for universal domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6258–6267.

Dempster, A. P. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2): 205–232.

Dempster, A. P. 2008. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, 57–72.

Fu, W.; Chen, Y.; Liu, W.; Yue, X.; and Ma, C. 2023. Evidence reconciled neural network for out-of-distribution detection in medical images. In *International conference on medical image computing and computer-assisted intervention*, 305–315.

Gao, Z.; Hu, D.; Jiang, X.; Lu, H.; Shen, H. T.; and Xu, X. 2024a. Enhanced Experts with Uncertainty-Aware Routing for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9650–9659.

Gao, Z.; Jiang, X.; Xu, X.; Shen, F.; Li, Y.; and Shen, H. T. 2024b. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26876–26885.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guo, Z.; Jin, T.; Chen, J.; and Zhao, Z. 2024. Classifier-guided gradient modulation for enhanced multimodal learning. *Advances in Neural Information Processing Systems*, 37: 133328–133344.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2551–2566.

Hu, D.; Jiang, X.; Sun, Z.; Shen, F.; and Xu, X. 2025a. Heterogeneous Graph Embedding for Multimodal Multi-Label Emotion Recognition. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 460–468.

Hu, D.; Jiang, X.; Sun, Z.; Yang, H.; Peng, C.; Yan, P.; Shen, H. T.; and Xu, X. 2025b. Geometric Gradient Divergence Modulation for Imbalanced Multimodal Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1337–1345.

Hu, Y.; Ou, Y.; Zhao, X.; Cho, J.-H.; and Chen, F. 2021. Multidimensional uncertainty-aware evidential neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 7815–7822.

Huang, H.; Qin, C.; Liu, Z.; Ma, K.; Chen, J.; Fang, H.; Ban, C.; Sun, H.; and He, Z. 2025. Trusted unified feature-neighborhood dynamics for multi-view classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17413–17421.

Jiang, X.; Huang, Z.; Xu, X.; Song, J.; Shen, F.; and Shen, H. T. 2025. PHGC: Procedural Heterogeneous Graph Completion for Natural Language Task Verification in Egocentric Videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8615–8624.

Jiang, X.; Xu, X.; Lu, H.; He, L.; and Shen, H. T. 2024a. Joint objective and subjective fuzziness denoising for multimodal sentiment analysis. *IEEE Transactions on Fuzzy Systems*, 33(1): 15–27.

Jiang, X.; Xu, X.; Zhu, L.; Sun, Z.; Cichocki, A.; and Shen, H. T. 2024b. Resisting noise in pseudo labels: Audible video event parsing with evidential learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Li, C.; Li, K.; Ou, Y.; Kaplan, L. M.; Jøsang, A.; Cho, J.-H.; Jeong, D. H.; and Chen, F. 2024. Hyper evidential deep learning to quantify composite classification uncertainty. *arXiv preprint arXiv:2404.10980*.

Lin, R.; and Hu, H. 2023. Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26: 2740–2755.

Liu, S.; Li, X.; Chen, Y.; Jiang, Y.; and Cong, G. 2025. Disentangling Dynamics: Advanced, Scalable and Explainable Imputation for Multivariate Time Series. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, W.; Chen, Y.; and Yue, X. 2025. Enhancing Multi-View Classification Reliability with Adaptive Rejection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18969–18977.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Niu, T.; Zhu, S.; Pang, L.; and El Saddik, A. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*, 15–27.

Pandey, D. S.; and Yu, Q. 2023. Evidential conditional neural processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9389–9397.

Qu, J.; Chen, Y.; Yue, X.; Fu, W.; and Huang, Q. 2024. Hyper-opinion evidential deep learning for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37: 84645–84668.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 746–760.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 567–576.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Wang, Y.; Liu, L.; Yuan, C.; Li, M.; and Liu, J. 2024. Negative-sensitive framework with semantic enhancement for composed image retrieval. *IEEE Transactions on Multimedia*, 26: 7608–7621.

Wang, Z.; Xu, X.; Zhu, L.; Bin, Y.; Wang, G.; Yang, Y.; and Shen, H. T. 2025. Evidence-Based Multi-Feature Fusion for Adversarial Robustness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wei, S.; Luo, Y.; Wang, Y.; and Luo, C. 2024. Robust multimodal learning via representation decoupling. In *European Conference on Computer Vision*, 38–54.

Wei, Y.; and Hu, D. 2024. Mmpareto: boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*.

Xu, C.; Si, J.; Guan, Z.; Zhao, W.; Wu, Y.; and Gao, X. 2024. Reliable conflictive multi-view learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 16129–16137.

Xu, S.; Sun, Y.; Li, X.; Duan, S.; Ren, Z.; Liu, Z.; and Peng, D. 2025. Noisy label calibration for multi-view classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21797–21805.

Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7617–7630.

Yang, Z.; Wei, Y.; Liang, C.; and Hu, D. 2024. Quantifying and enhancing multi-modal robustness with modality preference. *arXiv preprint arXiv:2402.06244*.

Zhang, Q.; Wu, H.; Zhang, C.; Hu, Q.; Fu, H.; Zhou, J. T.; and Peng, X. 2023. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, 41753–41769.

Zhao, X.; Chen, F.; Hu, S.; and Cho, J.-H. 2020. Uncertainty aware semi-supervised learning on graph data. *Advances in neural information processing systems*, 33: 12827–12836.