QUANTIFYING AND ENHANCING MULTI-MODAL ROBUSTNESS WITH MODALITY PREFERENCE

Zequn Yang[1], **Yake Wei**[1], **Ce Liang**[1], **Di Hu**[1]*
Gaoling School of Artificial Intelligence, Renmin University of China[1]
{zqyang,yakewei,liangce158,dihu}@ruc.edu.cn

ABSTRACT

Multi-modal models have shown a promising capability to effectively integrate information from various sources, yet meanwhile, they are found vulnerable to pervasive perturbations, such as uni-modal attacks and missing conditions. To counter these perturbations, robust multi-modal representations are highly expected, which are positioned well away from the discriminative multi-modal decision boundary. In this paper, different from conventional empirical studies, we focus on a commonly used joint multi-modal framework and theoretically discover that larger uni-modal representation margins and more reliable integration for modalities are essential components for achieving higher robustness. This discovery can further explain the limitation of multi-modal robustness and the phenomenon that multi-modal models are often vulnerable to attacks on the specific modality. Moreover, our analysis reveals how the widespread issue, that the model has different preferences for modalities, limits the multi-modal robustness by influencing the essential components and could lead to attacks on the specific modality highly effective. Inspired by our theoretical finding, we introduce a training procedure called *Certifiable Robust Multi-modal Training* (CRMT), which can alleviate this influence from modality preference and explicitly regulate essential components to significantly improve robustness in a certifiable manner. Our method demonstrates substantial improvements in performance and robustness compared with existing methods. Furthermore, our training procedure can be easily extended to enhance other robust training strategies, highlighting its credibility and flexibility.

1 INTRODUCTION

As data are often presented from different perspectives, like text, images, and audio, how to effectively exploit and integrate information from multiple sources becomes important. This has given rise to the concept of multi-modal learning, which serves as a potent approach that enables a more comprehensive understanding of complex concepts and facilitates more effective knowledge acquisition for different sources of information (Wei et al., 2022) Nowadays, multi-modal learning has demonstrated its remarkable ability in various tasks, including scene understanding (Antol et al., 2015; Yang et al., 2022; Li et al., 2022a), and emotion recognition (Tripathi et al., 2018; Chudasama et al., 2022).

However, real-world data are often perturbed, such as attacks, and missing modality, which impacts the performance of multi-modal model (Kumar et al., 2020). As a result, multi-modal robustness, which refers to the model's ability to defend against such perturbations, has received increasing attention in recent studies (Bednarek et al., 2020; Vishwamitra et al., 2021). Unlike data from a single modality, multi-modal data can be perturbed across all modalities. Therefore, for robustness, multi-modal models should have the ability to resist attacks on both individual and multiple modalities. Nevertheless, experiments have suggested that multi-modal models could perform badly when encountering perturbation (Noever & Noever, 2021) or missing modality (Yu et al., 2020; Ma et al., 2022). Based on these observations about the vulnerability of the multi-modal model, how to achieve a robust multi-modal model arises and attaches more focus.

To address this, previous methods improve the training strategies to obtain a robust multi-modal model (Liang et al., 2021; Ding et al., 2021). Specifically, certain studies extend uni-modal robust training strategies, like adversarial training and mixup, to learn a discriminative multi-modal decision boundary (Li et al., 2022b; Maheshwari et al., 2023). Others take steps like cross-modal

*Corresponding author

alignment (Tian & Xu, 2021) to enhance the connection among modalities and obtain compact and robust multi-modal representation. However, even if they empirically improve robustness to some degree, they still lack in-depth theoretical analysis to understand the resilience of multi-modal models to perturbations, which is vital for safety-critical applications. More importantly, a universal phenomenon can be observed for these robust training methods that the adversarial attack on specific modalities could be more effective than others. As shown in Figure 1, the ℓ_2-PGD attack is more effective on modality $\#a$ than modality $\#v$ for both Joint Training and three widely used robust training methods on the Kinetics Sounds dataset.

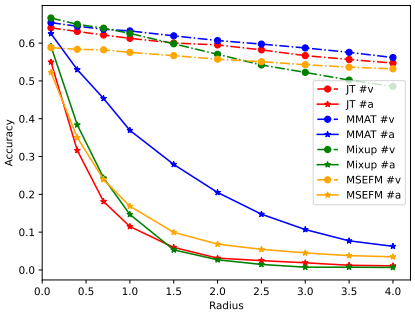


Figure 1: Accuracy of different multi-modal robust training methods compared with Joint Training (JT) baseline under ℓ_2-PGD attack with a range of radius for modality $\#v$ (vision) and $\#a$ (audio) respectively on Kinetics Sounds dataset. Results show that all these methods are more vulnerable to attacks on the specific modality $\#a$.

To deeply understand multi-modal robustness and elucidate this phenomenon, we first depict the multi-modal decision boundary by integrating uni-modal representation margins. Then, we derive a lower bound for the perturbation radius that the multi-modal model can consistently defend against. We discover that larger uni-modal margins coupled with reasonable integration are crucial for enhanced multi-modal robustness. With the above theoretical results, we investigate the pervasive issue (Wang et al., 2020; Peng et al., 2022) that multi-modal models exhibit a pronounced preference for a particular modality. This preferred modality profoundly influences the model's decision, then resulting in the above robustness phenomenon. On one hand, when a specific preferred modality is substantial enough to rely upon, the model will be reluctant to learn from other modalities, leading to the imbalance problem (Huang et al., 2022; Wu et al., 2022). This imbalance problem hinders the enhancement of the uni-modal margin on the other modality, thus limiting the robustness. On the other hand, since the multi-modal model heavily relies on the preferred modality, the corresponding factor used in modality integration becomes larger, which will amplify the variation of the uni-modal margin in decision-making. Hence, in case the preferred modality is vulnerable, the multi-modal model becomes more vulnerable to multi-modal attack. Further, this preference for vulnerable modality makes attacks on the preferred modality significantly more effective than other ones, explaining the observation in Figure 1.

Since the essential components of robustness have an interrelation with each other, directly applying regulation is hard to guarantee higher robustness. To address this, we employ an orthogonal-based framework that formulates an alternative bound, which eliminates the interrelation and explicitly presents the integration. Building upon our theoretical analysis, we introduce a two-step *Certifiable Robust Multi-modal Training* (CRMT) procedure to ensure progressively superior robustness. Initially, we redefine uni-modal margins related to the reliability of the modality in this framework. Then we propose to regulate the unreliable modality by enlarging its margin, which can alleviate the imbalanced problem brought by modality preference. Further, our approach adjusts the integration of modalities considering the improvement of certified bound. These steps not only mitigate the large gap between robustness against attack on each modality but also credibly guarantee higher multi-modal robustness. To validate our method, we conduct extensive experiments to present the advanced robustness against both uni-modal and multi-modal attacks. Our contributions are as follows:

1. We focus on a commonly used multi-modal model, offering invaluable insights into the essential components influencing multi-modal robustness.
2. We present analyses highlighting how multi-modal preference limits the multi-modal robustness and contributes to the vulnerability of multi-modal models towards specific modalities.
3. Drawing from our theoretical findings, we introduce a two-step training procedure, alleviating the limitation brought by modality preference. Our method can effectively enhance both performance and robustness over three real-world multi-modal datasets.

2 Related Work

Multi-modal Robustness Analysis. Recent studies have highlighted the vulnerability of deep neural networks (DNNs) to attacks and perturbations (Goodfellow et al., 2014; Madry et al., 2017),

raising significant concerns regarding their deployment. With the presence of multiple modalities, the forms of perturbations include uni-modal attacks, multi-modal attacks (Schlarmann & Hein, 2023), and modality missing (Lee et al., 2023). One possible way to enable the models to resist these perturbations is to design robust training strategies to obtain a reliable decision boundary that is distanced from samples (Li et al., 2021). Certain methods, such as multi-modal adversarial training (Li et al., 2022b) and incorporating uni-modal tasks (Ma et al., 2022), have proven pertinent in this context. On the other hand, studies focus on the latent consistency among modalities and propose to enhance modality interaction like alignment (Tsai et al., 2018) to obtain compact and robust representation (Bednarek et al., 2020). Among these empirical works, we identify a universal phenomenon that multi-modal models are commonly vulnerable to a certain modality (Liang et al., 2021), which currently lacks a comprehensive theoretical explanation. This motivates us to theoretically figure out the essential components determining the multi-modal robustness and explain the vulnerable modality phenomenon.

Multi-modal imbalance problem. Multi-modal learning is a significant approach for achieving a richer understanding across various tasks (Baltrušaitis et al., 2018; Jiang et al., 2021). However, even with the potential for a more comprehensive latent representation from multi-modal models (Huang et al., 2021), their performance might not always surpass that of the best uni-modal counterparts (Wang et al., 2020; Huang et al., 2022). Due to the inherent differences between modalities, a multi-modal model might prefer or lean towards a modality that is easier to learn, potentially overlooking others (Wu et al., 2022). In response, various strategies are proposed to strengthen the learning of uni-modalities, thereby enhancing the generalization of multi-modal model (Wang et al., 2020; Peng et al., 2022; Fan et al., 2023; Xu et al., 2023b). However, these researches do not elucidate how this modality preference problem impacts robustness against perturbations, which is our main focus.

Certified robustness. To describe the robustness of the model, certified robustness is introduced to describe the size of permissible perturbations that a model can consistently defend against. Techniques such as randomized smoothing (Cohen et al., 2019; Rosenfeld et al., 2020; Yang et al., 2021; Salman et al., 2019) and interval bound propagation (Zhang et al., 2018; Lyu et al., 2021; Xu et al., 2020) are widely utilized to relax the model, which assists in determining the certificate bounds. Additionally, the Lipschitz constant can serve as an intuitive metric to illustrate how perturbations influence the decision of models (Weng et al., 2018; Leino et al., 2021). However, these methods are tailored for inputs with only a single modality but fail to focus on the primary concerns with multiple modalities. In our research, we establish certified robustness for multi-modal models and determine the essential components influenced by modality preference that limit the robustness.

3 METHOD

3.1 PRELIMINARIES

Multi-modal Framework. We consider a general K-way classification problem with a vectorized input sample $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \mathbb{R}^{d_1+d_2}$ consisting of two modalities, and a ground truth label $y \in [K]$. We consider the commonly used joint learning framework in multi-modal learning, where all of the modalities are projected into the shared space for downstream tasks (Huang et al., 2022). Concretely, the uni-modal representations are extracted by encoders $\phi^{(m)}$ and concatenated to form the joint representation, which is then mapped to the output space using a linear classifier, where $W \in \mathbb{R}^{K\times(\dim(\phi^{(1)})+\dim(\phi^{(2)}))}$ and $\boldsymbol{b} \in \mathbb{R}^K$ are the weight matrix and bias respectively, and $\dim(\cdot)$ represents the dimension. The logits output of the multi-modal model can be denoted as $h(\boldsymbol{x}) = W[\phi^{(1)}(\boldsymbol{x}^{(1)}); \phi^{(2)}(\boldsymbol{x}^{(2)})] + \boldsymbol{b} = W^{(1)}\phi^{(1)}(\boldsymbol{x}^{(1)}) + W^{(2)}\phi^{(2)}(\boldsymbol{x}^{(2)}) + \boldsymbol{b}$, where $W^{(m)} \in \mathbb{R}^{K\times\dim(\phi^{(m)})}$ is the part of classifier W related to the m-th modality.

Multi-modal robustness. To assess the certified robustness of a multi-modal model, we can measure the radius of the smallest perturbation that shifts a sample to the decision boundary. In other words, any perturbation that falls within this radius can be defended against.

Definition 3.1. Suppose the multi-modal model h can correctly classify the sample $\boldsymbol{x}$, *i.e.* $\forall k \neq y, h_y(\boldsymbol{x}) > h_k(\boldsymbol{x})$, where h_k denotes the logit score of the k-th class. The robustness radius of the multi-modal model towards the sample $\boldsymbol{x}$ can be defined as:

$$P(\boldsymbol{x}) = \min_{\boldsymbol{x}'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \qquad s.t.\ \exists j \neq y, h_y(\boldsymbol{x}') = h_j(\boldsymbol{x}'). \tag{1}$$

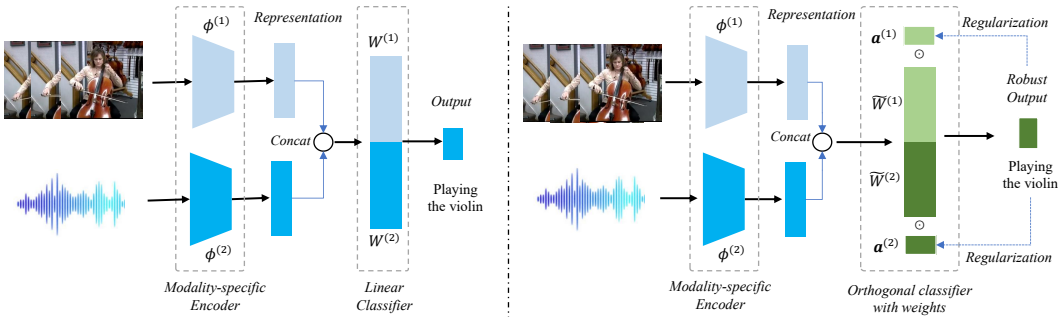


Figure 2: Illustration of traditional multi-modal joint learning framework (Baltrušaitis et al., 2018) (left) and our framework introducing orthogonality into each uni-modal classifier (right). Our framework can be easily applied to explicit regularization to achieve larger certified robustness.

Equation 1 seeks to identify the smallest perturbation, denoted as $\boldsymbol{x} - \boldsymbol{x}'$, that results in reaching the decision boundary between the ground truth y and its nearest class j. Hence, any smaller perturbation can always be defended against. Here, we use ℓ_2-norm to measure the radius, reflecting the overall size of the perturbation (Tsuzuku et al., 2018; Carlini & Wagner, 2017).

3.2 Certified Robustness for Multi-modal Model

In this section, we endeavor to uncover the essential components that impact multi-modal robustness. Referring to Figure 1, there is a notable variation in the effectiveness of perturbations on different modalities. This observation prompts us to distinguish the differences within each uni-modality. Therefore, we introduce the concept of *uni-modal margin*, which quantifies the distance between a uni-modal representation and the uni-modal decision boundary. To elaborate further, $(W_{y\cdot}^{(m)} - W_{k\cdot}^{(m)})\phi(\boldsymbol{x}^{(m)}) = 0$ signifies that sample $\boldsymbol{x}^{(m)}$ is positioned on the decision boundary between classes y and k for m-th modality. The uni-modal representation margin can be defined as:

Definition 3.2. Given the uni-modal encoder $\phi^{(m)}$ and the classifier $W^{(m)}$, the margin on representation space between ground truth y and other label k is defined as:

$$\zeta_k^{(m)}(\boldsymbol{x}^{(m)}) = \frac{(W_{y\cdot}^{(m)} - W_{k\cdot}^{(m)})\phi^{(m)}(\boldsymbol{x}^{(m)})}{\left\| W_{y\cdot}^{(m)} - W_{k\cdot}^{(m)} \right\|_2}. \tag{2}$$

In this context, a larger margin indicates the uni-modality is more reliable in distinguishing these two classes. Using this margin definition, we can re-examine the constraint conditions in Equation 1, which outline the perturbed sample located on the multi-modal decision boundary where y and the closest class j are tied in the output space:

$$h_y(\boldsymbol{x}') - h_j(\boldsymbol{x}') = c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}'^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}'^{(2)}) + \beta_j = 0. \tag{3}$$

Inspired by Equation 3, we observe that the multi-modal decision boundary can be described as integration of different uni-modal margins of the perturbed sample with factors $c_j^{(m)} = \|W_{y\cdot}^{(m)} - W_{j\cdot}^{(m)}\|_2$ and constant term $\beta_j = b_y - b_j$. The integration factors $c_j^{(m)}$ quantify how variations in the uni-modal margin influence multi-modal decisions. When perturbing a sample $\boldsymbol{x}$ towards the decision boundary, a larger factor $c_j^{(m)}$ indicates altering the margin of the corresponding modality is more effective. Meanwhile, when the sample is perturbed, how the uni-modal margin varies also depends on the size of the perturbation. Thus, we propose to introduce the Lipschitz constant $\tau_j^{(m)}$ for the uni-modal margin. This Lipschitz constant is the smallest constant to limit the local variation range of a certain function (Finlay et al., 2018), which is given by:

$$|\zeta_j^{(m)}(\boldsymbol{x}^{(m)}) - \zeta_j^{(m)}(\boldsymbol{x}'^{(m)})| \leq \tau_j^{(m)} \left\| \boldsymbol{x}^{(m)} - \boldsymbol{x}'^{(m)} \right\|_2. \tag{4}$$

Then, we can provide the lower bound for the perturbation radius for multi-modal robustness.

Theorem 3.3. *Given an input $\boldsymbol{x}$ with ground-truth label $y \in [K]$ and the closest label $j \neq y$, $\zeta_j^{(m)}(\boldsymbol{x}^{(m)})$ as the representation margin for m-th modality with Lipschitz constraint $\tau_j^{(m)}$, and the*

integration factor $c_j^{(m)}$. Define $\boldsymbol{x}'$ as the perturbed sample, and $\boldsymbol{x} - \boldsymbol{x}'$ as the perturbation. The lower bound for the perturbation radius can be described as:

$$P(\boldsymbol{x}) = \min_{\boldsymbol{x}'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \geq \frac{c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j}{\sqrt{(c_j^{(1)}\tau_j^{(1)})^2 + (c_j^{(2)}\tau_j^{(2)})^2}}$$
$$where \quad j \neq y \quad s.t. \quad c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}'^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}'^{(2)}) + \beta_j = 0. \tag{5}$$

The detailed proof can be found in the Appendix. Based on the certified bound above, we deduce that the multi-modal robustness depends on three primary factors: uni-modal representation margins $\zeta_j^{(m)}$, integration $c_j^{(m)}$, and bias difference β_j. Firstly, robustness increases proportionally with the enhancement of the uni-modal representation margin $\zeta_j^{(m)}$. Secondly, the integration $c_j^{(m)}$ is related to both uni-modal margins and the uni-modal Lipschitz constant, the expected integration for robustness requires considering both. Thus, a reasonable choice of integration factor can benefit higher robustness. Thirdly, when the sample is perturbed, the bias difference term β_j stays invariant, since it depends on class y and j rather than the specific sample. Based on this fact, the bias difference is not considered in our analysis about model robustness. In a nutshell, we recognize that uni-modal margins and the integration of modalities are two essential components. In the following section, we will analyze how these essential components vary and influence the robustness of multi-modal models, especially under modality preference.

3.3 Analysis about modality preference for robustness

Since modalities have different amounts of information, some specific modalities might hold more significance than others in decision-making processes (Gat et al., 2021). As a result, it is widely recognized that multi-modal models tend to show a preference for and predominantly rely on specific modality (Huang et al., 2022; Wu et al., 2022). However, this preference or over-reliance on the specific modality poses challenges for achieving a robust multi-modal model. We delve deeper into how such preferences impact the two essential components of the certified bound in Equation 5.

Uni-modal representation margin. As shown in Equation 5, a larger uni-modal margin $\zeta_j^{(m)}(\boldsymbol{x}^{(m)})$ determines higher certified robustness. However, when the learned information in the preferred modality is sufficiently reliable, the multi-modal model is reluctant to learn more information from other modalities (Wu et al., 2022; Huang et al., 2022). Thus, the modality preference leads to an imbalance problem hindering the development of uni-modal representations, resulting in a narrower representation margin and ultimately constraining the certified robustness of the multi-modal model.

Integration of modalities. According to modality preference, the decision of the multi-modal model highly depends on a specific modality. Thus, considering the integration of modalities, the preferred modality contributes more and is allocated a larger integration factor $c_j^{(m)}$, which could amplify the variation of the uni-modal margin in multi-modal decision-making. Since the preference is only determined by whether the modality with ideal discriminative ability, the multi-modal model could prefer a vulnerable modality, which has a larger $\tau_j^{(m)}$. Thus, the perturbation for this preferred but vulnerable modality leads to larger variations in multi-modal margins, which is further amplified in decision-making. Motivated by this phenomenon, we define $\eta^{(m)} = c_j^{(m)}\tau_j^{(m)}$ as the vulnerability indicator of modality m. When the multi-modal model exhibits a preference for a vulnerable modality 1, there is a significant imbalance in this indicator, with $\eta^{(1)} \gg \eta^{(2)}$. Consequently, we observe:

$$P(\boldsymbol{x}) \geq \frac{c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j}{\sqrt{(\eta^{(1)})^2 + (\eta^{(2)})^2}} \approx \frac{c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j}{\eta^{(1)}}. \tag{6}$$

That is to say, the modality preference on vulnerable modality leaves an unreasonable imbalance on $\eta^{(m)}$, thus the multi-modal robustness is highly dependent on the modality with a larger vulnerability indicator. In this way, distinctly attacking the vulnerable modalities alone is enough to obscure the model. Here we further provide the multi-modal robustness under uni-modal attack case:

Proposition 3.4. *Following the setting in Theorem 3.3,* w.l.o.g. *considering the perturbation on* 1*-th modality* $\boldsymbol{x}'^{(1)}$*, the lower bound for the uni-modal perturbation radius can be described as:*

$$\min_{\boldsymbol{x}'^{(1)}} \left\| \boldsymbol{x}^{(1)} - \boldsymbol{x}'^{(1)} \right\|_2 \geq \frac{c_j^{(1)} \zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)} \zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j}{c_j^{(1)} \tau_j^{(1)}} \quad \text{where} \quad j \neq y \quad s.t. \quad c_j^{(1)} \zeta_j^{(1)}(\boldsymbol{x}'^{(1)}) + c_j^{(2)} \zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j = 0. \tag{7}$$

The lower bounds of different uni-modal perturbations share identical numerators but differ in denominators, the vulnerability indicator $\eta^{(m)}$. The larger indicator on preferred modality reduces the lower bound for perturbations on this modality, thus making attacks on this preferred modality more effective, explaining the observation in Figure 1.

3.4 Certifiable Robust Multi-modal Training

Based on the above analysis, we target to improve the uni-modal representation margin $\zeta_j^{(m)}$ and adjust the integration of modalities $c_j^{(m)}$ as regulation for higher certified robustness. However, these regulations are intricately linked with the linear classifier W, potentially leading to conflicts in optimization objectives. As a result, stably enhancing the certified robustness presents challenges. Additionally, determining the Lipschitz constant for the margin is computationally demanding. To address these challenges, we propose to adopt orthogonality (Huang et al., 2018) within uni-modal linear classifiers as our framework. Detailedly, we ensure that in each modality, the class-specific vectors $\tilde{W}_{k\cdot}^{(m)}, k \in [K]$ are unit and orthogonal. Since the valuable information between modalities is different, we apply the weight $\boldsymbol{a}^{(m)} \in \mathbb{R}^K$ to lead the model focusing on more reliable modalities. Integrating these weights can enhance the model's ability to effectively utilize the valuable information from each modality. Therefore, learning of uni-modal representations and integration of modalities can be decoupled. The score corresponding to the k-th class can be expressed as:

$$\tilde{h}_k(\boldsymbol{x}) = a_k^{(1)} \tilde{W}_{k\cdot}^{(1)} \phi^{(1)}(\boldsymbol{x}^{(1)}) + a_k^{(2)} \tilde{W}_{k\cdot}^{(2)} \phi^{(2)}(\boldsymbol{x}^{(2)}) + \tilde{b}_k, \tag{8}$$

where $\tilde{W}^{(m)} \in \mathbb{R}^{K \times \dim(\phi^{(m)})}$ is the matrix with orthogonal rows, satisfying $\tilde{W}^{(m)}(\tilde{W}^{(m)})^T = I_K$. We can use $\tilde{W}_{k\cdot}^{(m)} \Phi(\boldsymbol{x}^{(m)})$ as the uni-modal score, which represents whether the uni-modal representation is well learned toward the k-th class. With this framework in place, we are able to express the new Lipschitz constant $\tilde{\tau}_k^{(m)}$ for the uni-modal score on class k in the m-th modality as:

$$|\tilde{W}_{k\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - \tilde{W}_{k\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}'^{(m)})| \leq \tilde{\tau}_k^{(m)} \left\| \boldsymbol{x}^{(m)} - \boldsymbol{x}'^{(m)} \right\|_2. \tag{9}$$

With these definitions, we can derive the certified bound for the multi-modal perturbation radius, which is distinctly tailored to the framework employing orthogonal classifiers:

Theorem 3.5. *Given an input* $\boldsymbol{x}$ *with ground-truth label* $y \in [K]$ *and the closest label* $j \neq y$*, the orthogonal classifier* $\tilde{W}^{(m)}$*, the modality-specific weight* $\boldsymbol{a}^{(m)}$*, the Lipschitz constant* $\tilde{\tau}_j^{(m)}$*, and the difference of the bias* $\tilde{\beta}_j = \tilde{b}_y - \tilde{b}_j$*. The lower bound for the perturbation radius with the orthogonal-based framework can be described as:*

$$P(\boldsymbol{x}) \geq \frac{\sum_{m=1}^{2} \left(a_y^{(m)} \tilde{W}_{y\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - a_j^{(m)} \tilde{W}_{j\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) \right) + \tilde{\beta}_j}{\sqrt{\sum_{m=1}^{2} \left(a_y^{(m)} \tilde{\tau}_y^{(m)} + a_j^{(m)} \tilde{\tau}_j^{(m)} \right)^2}} \quad \text{where} \quad j \neq y, \quad s.t. \quad \tilde{h}_y(\boldsymbol{x}') = \tilde{h}_j(\boldsymbol{x}'). \tag{10}$$

The proof positions in the Appendix. With this lower bound, we can discuss designing a regulation method to explicitly enhance the certified robustness. Firstly, as analyzed in Section 3.3, the imbalance problem brought by modality preference impacts the unreliable modal representation margin. To explicitly regulate the uni-modal encoder and classifier independent from integration, we redefine the margin as the difference in uni-modal score between ground truth and the other label, which can also reflect the reliability of uni-modality. Then we propose to enlarge the relatively small margin through regularization, which is expressed as follows:

$$\max_{\tilde{W}^{(m)}, \phi^{(m)}} \min_{m; k \neq y} \quad \tilde{W}_{y\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - \tilde{W}_{k\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}). \tag{11}$$

Table 1: Experimental results of adversarial accuracy results. Bold and underlined values indicate the top and the runner-up results. Combined with JT, MMAT, and Mixup, our proposed CRMT-based methods can improve both performance and robustness.

Attack	w\o			FGM			ℓ_2 PGD		
Datasets	KS	UCF101	VGGS	KS	UCF101	VGGS	KS	UCF101	VGGS
JT	0.643	0.742	0.496	0.288	0.506	0.157	0.268	0.431	0.056
GB	0.704	<u>0.784</u>	0.529	0.371	0.432	0.279	0.344	0.205	0.174
OGM	0.651	0.743	0.498	0.292	0.492	0.190	0.270	0.377	0.069
PMR	0.689	0.742	0.503	0.350	0.455	0.166	0.331	0.292	0.059
MSEFM	0.627	0.721	0.492	0.390	0.483	0.187	0.376	0.243	0.115
MMAT	0.656	0.728	0.509	<u>0.514</u>	<u>0.609</u>	<u>0.413</u>	<u>0.507</u>	<u>0.598</u>	<u>0.409</u>
Mixup	0.669	0.717	0.507	0.347	0.413	0.278	0.327	0.192	0.135
CMRT	<u>0.758</u>	**0.789**	0.526	0.491	0.515	0.248	0.468	0.433	0.124
CMRT-AT	**0.762**	0.759	**0.538**	**0.608**	**0.614**	**0.422**	**0.602**	**0.602**	**0.414**
CMRT-Mix	0.744	0.769	<u>0.537</u>	<u>0.514</u>	0.430	0.328	0.491	0.191	0.184

Based on Equation 11, we propose a regularization term, refining the maximum using the LogSumExp function to facilitate better optimization:

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{m=1}^{2}\frac{\sum_{k\neq y}\exp(\tilde{W}_{k\cdot}^{(m)}\phi^{(m)}(\boldsymbol{x}_i^{(m)}))}{\exp(\tilde{W}_{y\cdot}^{(m)}\phi^{(m)}(\boldsymbol{x}_i^{(m)}))}\right), \tag{12}$$

which is detailed in the Appendix. Furthermore, we can adjust the integration of different modalities by enlarging the lower bound in Equation 10. Subsequently, we propose a two-step training procedure called *Certifiable Robust Multi-modal Training* (CRMT), which can credibly obtain a robust multi-modal model. The training procedure of CRMT is as follows:

Step 1: optimize with cross-entropy loss and margin regularization with term ρ:
$\min_{\boldsymbol{a}^{(m)},\tilde{W}^{(m)},\phi^{(m)}} \rho L_1 + \frac{1}{N}\sum_{i=1}^{N} CE(h(\boldsymbol{x}_i), y_i)$, where CE is the cross-entropy loss function.

Step 2: fix $\tilde{W}^{(m)}, \phi^{(m)}$, update $\boldsymbol{a}^{(m)}$ to approach higher certified robustness:
$\min_{\boldsymbol{a}^{(m)}} \quad L_2 = -\frac{1}{N}\sum_{i=1}^{N} r(\boldsymbol{x}_i)$, where $r(\boldsymbol{x})$ is the lower bound in Equation 10.

As shown in the Appendix, we further demonstrate that only one iteration of our method can achieve considerable robustness, without a huge consumption of training cost.

4 Experiments

4.1 Setups

Dataset. We evaluate our method on different datasets including Kinetics-Sounds (Audio + Vision) (Arandjelovic & Zisserman, 2017), UCF101 (Optical flow + RGB) (Soomro et al., 2012), and VGGSound (Audio + Vision) (Chen et al., 2020). We use the backbone ResNet18 (He et al., 2016) as the encoder for each uni-modality. Details about these datasets are presented in the Appendix.

Multi-modal models. In this study, the comparison methods are selected to improve multi-modal learning and be suitable for the multi-modal joint training strategy. Comparison methods can be divided into two distinct groups: Firstly, methods address the imbalance problem caused by modality preference: *Gradient Blending* (GB) (Wang et al., 2020), *On-the-fly Gradient Modulation with generalization enhancement* (OGM) (Peng et al., 2022), *Prototypical Modal Rebalance* (PMR) (Fan et al., 2023). Secondly, methods aim at improving multi-modal robustness: *Multi-Modal Adversarial Training* (MMAT) (Li et al., 2022b), *Multi-modal mixup* (Mixup) (Madry et al., 2017; Li et al., 2022b), *MinSim+ExFMem* (MSEFM) (Tian & Xu, 2021). Our method can be extended to different training strategies, denoted as *Certifiable Robust Multi-modal Training with Joint Training* (CRMT-JT), CRMT *with Adversarial Training* (CRMT-AT), and CRMT *with Mixup* (CRMT-Mix).

Attack methods. Following previous work (Tsuzuku et al., 2018; Singla et al., 2021), we select *Fast Gradient Method* (FGM) (Goodfellow et al., 2014) and ℓ_2 *Projected Gradient Descent* (ℓ_2-PGD) (Madry et al., 2017) as two attack methods with attack size $\epsilon = 0.5$, which are widely used

Table 2: Performance against distinct uni-modal attack methods on KS dataset.

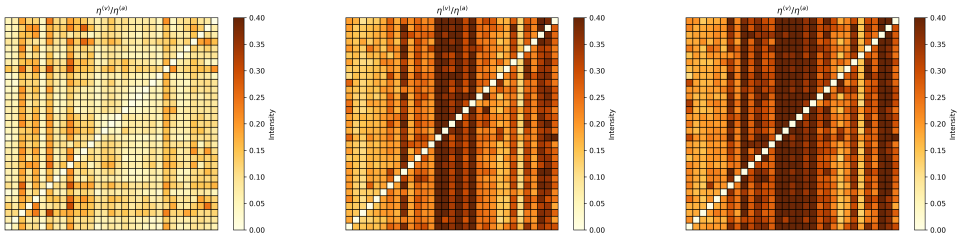
Method	Attack	w\o Clean	FGM #v	FGM #a	ℓ_2-PGD #v	ℓ_2-PGD #a	Missing modality #v	Missing modality #a
Baseline	JT	0.643	0.616	0.143	0.616	0.110	0.480	0.230
Imbalance	GB	0.704	0.658	0.195	0.656	0.157	0.483	0.430
	OGM	0.651	0.624	0.147	0.624	0.116	0.463	0.207
	PMR	0.689	0.649	0.181	0.649	0.152	0.484	0.315
Robustness	MSEFM	0.627	0.577	0.320	0.576	0.314	0.466	0.239
	MMAT	0.656	0.633	0.390	0.632	0.369	0.482	0.262
	Mixup	0.669	0.628	0.191	0.626	0.147	0.450	0.256
Ours	CRMT-JT	0.758	0.685	0.384	0.682	0.327	0.560	0.591
	CRMT-AT	**0.762**	0.703	**0.488**	0.698	**0.464**	0.547	**0.626**
	CRMT-Mix	0.744	**0.706**	0.377	**0.704**	0.320	**0.572**	0.566

for verifying multi-modal robustness with ℓ_2-norm. For uni-modal attack, we also introduce attacks FGM and ℓ_2-PGD with $\epsilon = 1.0$, and missing on uni-modality.

4.2 Robustness validation on multi-modal and uni-modal attack

Robustness against multi-modal attacks. We validate the robustness of our method under multi-modal attacks. Based on the experimental results presented in Table 1, we have identified four key observations that warrant attention. Firstly, while imbalance methods effectively enhance performance on clean samples, robustness methods can demonstrate more notable defense capabilities against various attacks. Secondly, the imbalance method GB can be superior to the robustness method MSEFM under some situations. That is because GB introduces a multi-task method to enhance the learning of uni-modality and improve the uni-modal margin, thus it can enhance the robustness methods according to our analysis. Thirdly, our proposed CRMT-based methods surpass the performance of the compared methods across these datasets in most cases. This superiority stems from our approach to addressing the imbalance problem through improving uni-modal representation margins, and the certificate adjustment of modality-specific weights for heightened certified robustness. Fourthly, the improvement of results for CRMT-AT and CRMT-Mix suggests that our training procedure can serve as a valuable component applicable to other robust training methods.

Robustness against distinct uni-modal attack. To substantiate the efficacy of our CRMT methods, we conducted additional experiments on the KS dataset, involving a broader range of different uni-modal attacks. As demonstrated in Table 2, the absence of modality #a leads to a larger performance decline than #v, since it is more preferred by multi-modal models. As shown in Table 2, previous multi-modal methods demonstrated poor performance when subjected to attacks on the preferred modality #a, aligning with our analysis. In contrast, our CRMT-based methods consider addressing the imbalance problem introduced by modality preference and adjusting integration, thus making the model more robust against uni-modal attacks. Furthermore, when encountering different uni-modal attacks, our CRMT-based approach shows superior performance and consistently ensures higher robustness in various scenarios. This strongly demonstrates the effectiveness and versatility of our proposed method in enhancing the robustness of multi-modal models.



(a) $\eta^{(v)}/\eta^{(a)}$ in MMAT. (b) $\eta^{(v)}/\eta^{(a)}$ in CRMT-AT step-1. (c) $\eta^{(v)}/\eta^{(a)}$ in CRMT-AT.

Figure 3: Evaluation of the ratio of vulnerability indicators between modality #v and #a (preferred). We illustrate the ratio in MMAT, CRMT-AT, and CRMT-AT with only the first training procedure.

Table 3: Extension results of adversarial accuracy with transformer-based method on VGGS dataset.

Attack	Clean	Uni-modal attack				Multi-modal attack	
Method	w\o	FGM #v	FGM #a	ℓ_2-PGD #v	ℓ_2-PGD #a	FGM	ℓ_2-PGD
MMT	0.465	0.433	0.259	0.431	0.228	0.331	0.326
CRMT-MMT	0.471	0.440	0.276	0.428	0.237	0.344	0.340

Imbalance on uni-modal vulnerability indicators. We verify the conclusion drawn in Section 3.3 that the imbalance problem introduced by modality preference leads to the vulnerability of attack on specific modality, as illustrated in Figure 1. To achieve this, we compare MMAT with our CRMT-AT approach. As depicted in Equation 7, $\eta^{(m)}$ is utilized as a measurement to assess the robustness of multi-modal models against uni-modal perturbations, and we employ the ratio of uni-modal vulnerability indicators $\eta^{(v)}/\eta^{(a)}$ (see Figure 3). It can be demonstrated that our approach diligently works to reduce the imbalance in the indicator η, effectively darkening the heated map. Furthermore, according to the robustness accuracy, our approach significantly reduces the disparity in attack performance between the two modalities (see Figure 4). This evidence serves to illustrate the efficacy of our method in mitigating the imbalance problem and ultimately enhancing multi-modal robustness.

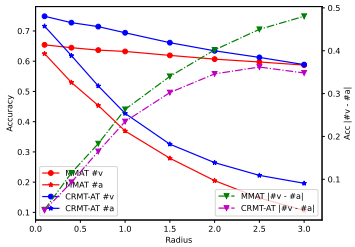


Figure 4: This figure presents the robustness accuracy against uni-modal attacks with different sizes, where the dotted line signifies the difference in robustness accuracy between two uni-modalities.

4.3 Validation for Effectiveness and Scalability

Ablation studies. To delve deeper into the efficacy of our training procedure, we conduct an ablation analysis, aiming to elucidate the contribution of each component of the proposed training procedure toward the overall result. When compared with the baseline Joint Training (JT), our results include the orthogonal-based framework and our proposed training procedure. We report our findings related to solely using the orthogonal framework (OJT), CRMT focusing exclusively on each step (CRMT-step-1 and CRMT-step-2), and our full proposal, CRMT-JT, which incorporates both steps. As depicted in Figure 5, it is evident that the orthogonal-based framework does not contribute to improved robustness. Meanwhile, both step-1 and step-2 of the training procedure can enhance robustness, particularly in the face of larger-size attacks. Furthermore, the results highlight that each step consistently enhances multi-modal robustness, making CRMT-JT superior in multi-modal robustness.

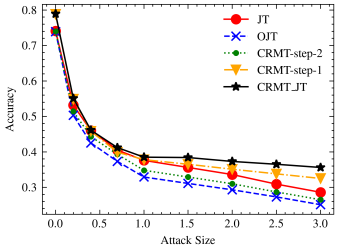


Figure 5: Ablation studies of our methods on the UCF101 dataset, revealing the effect of each part we introduced.

Extension studies to transformer. We further provide experiment results of our methods extended to the Multi-Modal Transformer-based framework with hierarchical attention (Xu et al., 2023a) (MMT) on the VGGS dataset. Both visual and audio class tokens are concatenated and linearly projected into the output space. We mainly evaluate the robustness under the Transformer model accurately, hence we train from scratch instead of using the pre-trained model. To validate our method, we only introduce the CRMT procedure step-1 on this transformer architecture as a comparison. That is because both audio and visual input influence each token, which is beyond our assumption. Based on the results in Table 3, our method can also improve the robustness of transformer-based architecture in most cases, which proves the broad prospects of our approach.

5 Conclusion

In this study, we present the essential components for multi-modal robustness and delve into the limitations imposed by modality preference. Additionally, we explore why multi-modal models exhibit vulnerabilities to attack specific modalities. Further, we introduce the Credible Multi-modal Robust Training procedure, a novel approach explicitly designed to boost the certified robustness.

6 ACKNOWLEDGEMENT

This research was supported by National Natural Science Foundation of China (NO.62106272), the Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), and Public Computing Cloud, Renmin University of China.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Michal Bednarek, Piotr Kicki, and Krzysztof Walas. On robustness of multi-modal fusion—robotics perspective. *Electronics*, 9(7):1152, 2020.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4652–4661, 2022.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Wenhao Ding, Baiming Chen, Bo Li, Kim Ji Eun, and Ding Zhao. Multimodal safety-critical scenarios generation for decision-making algorithms evaluation. *IEEE Robotics and Automation Letters*, 6(2):1551–1558, 2021.

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

Chris Finlay, Adam M Oberman, and Bilal Abbasi. Improved robustness to adversarial examples using lipschitz regularization of the loss. 2018.

Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34:21630–21643, 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). *arXiv preprint arXiv:2203.12221*, 2022.

Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Deepak Kumar, Chetan Kumar, Chun Wei Seah, Siyu Xia, and Ming Shao. Finding achilles' heel: Adversarial attack on multi-modal action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3829–3837, 2020.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952, 2023.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *International Conference on Machine Learning*, pp. 6212–6222. PMLR, 2021.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19108–19118, 2022a.

Juncheng B Li, Kaixin Ma, Shuhui Qu, Po-Yao Huang, and Florian Metze. Audio-visual event recognition through the lens of adversary. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 616–620. IEEE, 2021.

Juncheng B Li, Shuhui Qu, Xinjian Li, Po-Yao Bernie Huang, and Florian Metze. On adversarial robustness of large-scale audio visual learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 231–235. IEEE, 2022b.

Siting Li, Chenzhuang Du, Yue Zhao, Yu Huang, and Hang Zhao. What makes for robust multi-modal models in the face of missing modalities? *arXiv preprint arXiv:2310.06383*, 2023.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*, 2021.

Zhaoyang Lyu, Minghao Guo, Tong Wu, Guodong Xu, Kehuan Zhang, and Dahua Lin. Towards evaluating and training verifiably robust neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4308–4317, 2021.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing modality robustness in semi-supervised multi-modal semantic segmentation. *arXiv preprint arXiv:2304.10756*, 2023.

Frank Nielsen and Ke Sun. Guaranteed bounds on the kullback–leibler divergence of univariate mixtures. *IEEE Signal Processing Letters*, 23(11):1543–1546, 2016.

David A Noever and Samantha E Miller Noever. Reading isn't believing: Adversarial attacks on multi-modal neurons. *arXiv preprint arXiv:2103.10480*, 2021.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pp. 8230–8241. PMLR, 2020.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3677–3685, 2023.

Sahil Singla, Surbhi Singla, and Soheil Feizi. Improved deterministic l2 robustness on cifar-10 and cifar-100. *arXiv preprint arXiv:2108.04062*, 2021.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5601–5611, 2021.

Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Nishant Vishwamitra, Hongxin Hu, Ziming Zhao, Long Cheng, and Feng Luo. Understanding and measuring robustness of multimodal learning. *arXiv preprint arXiv:2112.12792*, 2021.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pp. 5276–5285. PMLR, 2018.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 33:1129–1141, 2020.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3480–3491, 2022.

Zhuolin Yang, Linyi Li, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, and Bo Li. On the certified robustness for ensemble models and beyond. *arXiv preprint arXiv:2107.10873*, 2021.

Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning. *arXiv preprint arXiv:2005.10987*, 2020.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *Advances in neural information processing systems*, 31, 2018.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5005–5013, 2022.

7 DETAILED ANALYSIS AND PROOFS IN SECTION 3

7.1 PROOF OF THEOREM 1

In this section, we provide proof for two lower bounds in Equation 5 and Equation 10.

Theorem 7.1. *Given an input $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ with ground-truth label $y \in [K]$ and the runner-up label $j \neq y$, $\zeta_j^{(m)}(\boldsymbol{x}^{(m)})$ as the representation margin for m-th modality, satisfying the Lipschitz constraint $\tau_j^{(m)}$, and the fusion factor $c_j^{(m)}$. Define $\boldsymbol{x}' = (\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)})$ as the perturbed sample, and $\boldsymbol{x} - \boldsymbol{x}'$ denotes the perturbation. The bound for the perturbation radius can be described as:*

$$
\begin{aligned}
P(\boldsymbol{x}) = \min_{\boldsymbol{x}'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \geq \frac{c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j}{\sqrt{(c_j^{(1)}\tau_j^{(1)})^2 + (c_j^{(2)}\tau_j^{(2)})^2}} \\
s.t.\ \ c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}'^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}'^{(2)}) + \beta_j = 0.
\end{aligned}
\tag{13}
$$

Proof. The key to this proof is how to bridge the gap between multi-modal perturbation and the uni-modal representation, by properly introducing the perturbed sample through the condition in Inequality 13 and the margin on perturbation $\zeta_j^{(m)}(\boldsymbol{x}'^{(m)})$. First, considering this constraint, we can obtain the following equation:

$$
\begin{aligned}
h_y(\boldsymbol{x}) - h_j(\boldsymbol{x}) &= c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j \\
&= c_j^{(1)}(\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) - \zeta_j^{(1)}(\boldsymbol{x}'^{(1)})) + c_j^{(2)}(\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) - \zeta_j^{(2)}(\boldsymbol{x}'^{(2)})),
\end{aligned}
\tag{14}
$$

where the formulation of the difference of margin occurs, which can serve as a bond to connect with the perturbation. Then consider the following inequality:

$$
\begin{aligned}
&c_j^{(1)}|\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) - \zeta_j^{(1)}(\boldsymbol{x}'^{(1)})| + c_j^{(2)}|\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) - \zeta_j^{(2)}(\boldsymbol{x}'^{(2)})| \\
&\leq c_j^{(1)}\tau_j^{(1)} \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}'^{(1)}\right\|_2 + c_j^{(2)}\tau_j^{(2)} \left\|\boldsymbol{x}^{(2)} - \boldsymbol{x}'^{(2)}\right\|_2 .
\end{aligned}
\tag{15}
$$

Once the condition $\forall m = 1, 2, \zeta_j^{(m)}(\boldsymbol{x}^{(m)}) \geq \zeta_j^{(m)}(\boldsymbol{x}'^{(m)})$ is satisfied, which imposes the minimum requirement on $\boldsymbol{x}'$, as proven later, we can establish the relationship with the perturbation. Furthermore, by employing the Cauchy Inequality, we obtain the following inequality:

$$
c_j^{(1)}\tau_j^{(1)} \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}'^{(1)}\right\|_2 + c_j^{(2)}\tau_j^{(2)} \left\|\boldsymbol{x}^{(2)} - \boldsymbol{x}'^{(2)}\right\|_2 \leq \sqrt{(c_j^{(1)}\tau_j^{(1)})^2 + (c_j^{(2)}\tau_j^{(2)})^2}\, \|\boldsymbol{x} - \boldsymbol{x}'\|_2 .
\tag{16}
$$

Thus, we have successfully established the bound for perturbation. However, two challenges remain unresolved. Firstly, we need to minimize the objective function by finding an optimal perturbation $\boldsymbol{x}'$. Secondly, the validity of condition $\zeta_j^{(m)}(\boldsymbol{x}^{(m)}) \geq \zeta_j^{(m)}(\boldsymbol{x}'^{(m)})$ cannot be guaranteed. In the subsequent sections, we will address these issues and provide formal proofs to support our claims.

Initially, drawing inspiration from the Inequality 7-16, we perform a straightforward transformation on the Inequality 13, aiming to establish a connection between multi-modal perturbation and uni-modal perturbation:

$$
\begin{aligned}
\min_{\boldsymbol{x}'} \sqrt{(c_j^{(1)}\tau_j^{(1)})^2 + (c_j^{(2)}\tau_j^{(2)})^2}\, \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \geq c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j \\
s.t.\ \ c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}'^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}'^{(2)}) + \beta_j = 0.
\end{aligned}
\tag{17}
$$

Subsequently, we arrive at the following inequality:

$$
\begin{aligned}
&\min_{\boldsymbol{x}'} \sqrt{(c_j^{(1)}\tau_j^{(1)})^2 + (c_j^{(2)}\tau_j^{(2)})^2}\, \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \\
\geq &\min_{\boldsymbol{x}'^{(1)},\boldsymbol{x}'^{(2)}} c_j^{(1)}\tau_j^{(1)} \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}'^{(1)}\right\|_2 + c_j^{(2)}\tau_j^{(2)} \left\|\boldsymbol{x}^{(2)} - \boldsymbol{x}'^{(2)}\right\|_2 \qquad \text{(Cauchy Inequality)} \\
\geq &\min_{\boldsymbol{x}'^{(1)},\boldsymbol{x}'^{(2)}} c_j^{(1)}|\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) - \zeta_j^{(1)}(\boldsymbol{x}'^{(1)})| + c_j^{(2)}|\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) - \zeta_j^{(2)}(\boldsymbol{x}'^{(2)})| \text{ (Lipschitz condition)} \\
= \quad & c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j, \qquad (*)
\end{aligned}
\tag{18}
$$

where the Equation $(*)$ is required to be proved. In the following, we prove this target equation in detail. The main concern is that whether $\zeta_j^{(m)}(\boldsymbol{x}^{(m)}) \geq \zeta_j^{(m)}(\boldsymbol{x}'^{(m)})$ holds for each modality m. As the robustness analysis focuses solely on correctly classified samples, where $\forall j \neq y, h_y(\boldsymbol{x}) > h_j(\boldsymbol{x})$, we can establish the following inequality:

$$
c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j > 0. \tag{19}
$$

The condition stated in Inequality 13 can be expressed as follows:

$$
h_y(\boldsymbol{x}') - h_j(\boldsymbol{x}') = c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}'^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}'^{(2)}) + \beta_j = 0. \tag{20}
$$

By incorporating Inequality 19 and Equation 20, we can obtain the following inequality:

$$
c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) > c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}'^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}'^{(2)}). \tag{21}
$$

Without loss of generality, we suppose $\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) < \zeta_j^{(1)}(\boldsymbol{x}'^{(1)})$, and we have:

$$
c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}'^{(2)}) + \beta_j < 0. \tag{22}
$$

Hence, since the margin function $\zeta_j^{(2)}$ is continuous, we define an additional variable to describe this process. Let $f(\lambda) = c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\lambda\boldsymbol{x}^{(2)} + (1-\lambda)\boldsymbol{x}'^{(2)}) + \beta_j$, we can observe that both $f(0)$ and $f(1)$ equal zero. Applying the Zero Existence Theorem, we can always find a $0 < \hat{\lambda} < 1$, and a perturbation $\hat{\boldsymbol{x}}'^{(2)} = \hat{\lambda}\boldsymbol{x}^{(2)} + (1-\hat{\lambda})\boldsymbol{x}'^{(2)}$. satisfying:

$$
c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\hat{\boldsymbol{x}}'^{(2)}) + \beta_j = 0. \tag{23}
$$

Hence, $\hat{\boldsymbol{x}} = (\boldsymbol{x}^{(1)}, \hat{\boldsymbol{x}}'^{(2)})$ is also a feasible perturbation satisfying the condition of Equation 13. And obviously,

$$
\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 = \|(1-\hat{\lambda})(\boldsymbol{x}^{(2)} - \boldsymbol{x}'^{(2)})\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}'\|_2, \tag{24}
$$

which contradicts the minimization constraint.

In conclusion, we have verify that $\zeta_j^{(m)}(\boldsymbol{x}^{(m)}) \geq \zeta_j^{(m)}(\boldsymbol{x}'^{(m)})$ for each modality m. Thus, we can approach the target equation in Equation 18 by the following equation:

$$
\begin{aligned}
&\min_{\boldsymbol{x}'^{(1)},\boldsymbol{x}'^{(2)}} c_j^{(1)}|\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) - \zeta_j^{(1)}(\boldsymbol{x}'^{(1)})| + c_j^{(2)}|\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) - \zeta_j^{(2)}(\boldsymbol{x}'^{(2)})| \\
= &\min_{\boldsymbol{x}'^{(1)},\boldsymbol{x}'^{(2)}} c_j^{(1)}(\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) - \zeta_j^{(1)}(\boldsymbol{x}'^{(1)})) + c_j^{(2)}(\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) - \zeta_j^{(2)}(\boldsymbol{x}'^{(2)})) \\
= &\, c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j.
\end{aligned}
\tag{25}
$$

Therefore, combined with Equation 18, we can obtain the target inequality.

$$
\min_{\boldsymbol{x}'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \geq \frac{c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(2)}) + \beta_j}{\sqrt{(c_j^{(1)}\tau_j^{(1)})^2 + (c_j^{(2)}\tau_j^{(2)})^2}}. \tag{26}
$$

□

7.2 PROOF OF THEOREM 2

Theorem 7.2. *Given an input $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ with ground-truth label $y \in [K]$ and the runner-up label $j \neq y$, the orthogonal classifier $\tilde{W}^{(m)}$, the modality-specific weight $\boldsymbol{a}^{(m)}$, the Lipschitz constant $\tilde{\tau}_j^{(m)}$, and the difference of the bias $\tilde{\beta}_j = \tilde{b}_y - \tilde{b}_j$. The bound for the perturbation radius can be described as:*

$$P(\boldsymbol{x}) \geq \frac{\sum_{m=1}^{2} \left(a_y^{(m)} \tilde{W}_{y\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - a_j^{(m)} \tilde{W}_{j\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) \right) + \tilde{\beta}_j}{\sqrt{\sum_{m=1}^{2} \left(a_y^{(m)} \tilde{\tau}_y^{(m)} + a_j^{(m)} \tilde{\tau}_j^{(m)} \right)^2}} \quad s.t.\ \exists j \neq y,\ \tilde{h}_y(\boldsymbol{x}') = \tilde{h}_j(\boldsymbol{x}'). \tag{27}$$

Proof. For simplify, here we use $\tilde{h}_y^{(m)}(\boldsymbol{x}^{(m)}) = a_y^{(m)} \tilde{W}_{y\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)})$ as the score of the sample $\boldsymbol{x}$ of y-th class in m-th modality. Hence, we have:

$$\tilde{h}_y = \tilde{h}_y^{(1)}(\boldsymbol{x}^{(1)}) + \tilde{h}_y^{(2)}(\boldsymbol{x}^{(2)}) + b_y. \tag{28}$$

Notice that the numerator term of the lower bound equals to $\tilde{h}_y(\boldsymbol{x}) - \tilde{h}_j(\boldsymbol{x})$, combining with the condition, we have:

$$\begin{aligned} \tilde{h}_y(\boldsymbol{x}) - \tilde{h}_j(\boldsymbol{x}) &= \tilde{h}_y(\boldsymbol{x}) - \tilde{h}_j(\boldsymbol{x}) - (\tilde{h}_y(\boldsymbol{x}') - \tilde{h}_j(\boldsymbol{x}')) \\ &= \sum_{m=1}^{2} \tilde{h}_y^{(m)}(\boldsymbol{x}^{(m)}) - \tilde{h}_j^{(m)}(\boldsymbol{x}^{(m)}) - (\tilde{h}_y^{(m)}(\boldsymbol{x}'^{(m)}) - \tilde{h}_j^{(m)}(\boldsymbol{x}'^{(m)})). \end{aligned} \tag{29}$$

Denote $\gamma_j^{(m)}(\boldsymbol{x}^{(m)}) = \tilde{h}_y^{(m)}(\boldsymbol{x}^{(m)}) - \tilde{h}_j^{(m)}(\boldsymbol{x}^{(m)})$. Inspired by the proof above, we have :

$$\gamma_j^{(m)}(\boldsymbol{x}^{(m)}) \geq \gamma_j^{(m)}(\boldsymbol{x}'^{(m)}) \tag{30}$$

holds for each modality m, which requires the minimization condition of $\boldsymbol{x}'$ and will be proved later. Thus, we can use the Lipschitz condition for j-th class:

$$|\tilde{W}_{j\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - \tilde{W}_{j\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}'^{(m)})| \leq \tilde{\tau}_j^{(m)} \left\| \boldsymbol{x}^{(m)} - \boldsymbol{x}'^{(m)} \right\|_2. \tag{31}$$

and we have:

$$\begin{aligned} &\sum_{m=1}^{2} |\gamma_j^{(m)}(\boldsymbol{x}^{(m)}) - \gamma_j^{(m)}(\boldsymbol{x}'^{(m)})| \\ =& \sum_{m=1}^{2} \left| \tilde{h}_y^{(m)}(\boldsymbol{x}^{(m)}) - \tilde{h}_y^{(m)}(\boldsymbol{x}'^{(m)}) - \tilde{h}_j^{(m)}(\boldsymbol{x}^{(m)}) + \tilde{h}_j^{(m)}(\boldsymbol{x}'^{(m)}) \right| \\ \leq& \sum_{m=1}^{2} \left| a_y^{(m)} \tilde{W}_{y\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - a_y^{(m)} \tilde{W}_{y\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}'^{(m)}) \right| \\ &\quad + \left| a_j^{(m)} \tilde{W}_{j\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - a_j^{(m)} \tilde{W}_{j\cdot}^{(m)} \phi^{(m)}(\boldsymbol{x}'^{(m)}) \right| \\ \leq& \left(a_y^{(1)} \tilde{\tau}_y^{(1)} + a_j^{(1)} \tilde{\tau}_j^{(1)} \right) \|\boldsymbol{x}^{(1)} - \boldsymbol{x}'^{(1)}\|_2 + \left(a_y^{(2)} \tilde{\tau}_y^{(2)} + a_j^{(2)} \tilde{\tau}_j^{(2)} \right) \|\boldsymbol{x}^{(2)} - \boldsymbol{x}'^{(2)}\|_2 \end{aligned} \tag{32}$$

Thus, we can further solve it by employing the Cauchy Inequality.

$$\sum_{m=1}^{2} |\gamma_j^{(m)}(\boldsymbol{x}^{(m)}) - \gamma_j^{(m)}(\boldsymbol{x}'^{(m)})| \leq \sqrt{\sum_{m=1}^{2} \left(a_y^{(m)} \tilde{\tau}_y^{(m)} + a_j^{(m)} \tilde{\tau}_j^{(m)} \right)^2} \|\boldsymbol{x} - \boldsymbol{x}'\|_2. \tag{33}$$

Further, we verify that when $\boldsymbol{x}'$ sets the size of perturbation the minimum value, Equation 30 satisfies. The proof is similar to the proof above. As our robustness analysis focuses solely on correctly classified samples, we can establish the following inequality:

$$\forall j \neq y, \tilde{h}_y(\boldsymbol{x}) - \tilde{h}_j(\boldsymbol{x}) = \gamma_j^{(1)}(\boldsymbol{x}^{(1)}) + \gamma_j^{(2)}(\boldsymbol{x}^{(2)}) > 0 \tag{34}$$

Without loss of generality, we suppose $\gamma_j^{(1)}(\boldsymbol{x}) < \gamma_j^{(1)}(\boldsymbol{x}')$. Considering the conditions above, we have:

$$\gamma_j^{(1)}(\boldsymbol{x}^{(1)}) + \gamma_j^{(2)}(\boldsymbol{x}'^{(2)}) < \gamma_j^{(1)}(\boldsymbol{x}'^{(1)}) + \gamma_j^{(2)}(\boldsymbol{x}'^{(2)}) = 0 \tag{35}$$

Hence, there exist a $\hat{\boldsymbol{x}} = (\boldsymbol{x}^{(1)}, \hat{\boldsymbol{x}}^{(2)})$, subject to :

$$\gamma_j^{(1)}(\boldsymbol{x}^{(1)}) + \gamma_j^{(2)}(\hat{\boldsymbol{x}}^{(2)}) = 0 \quad \text{and} \quad \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 < \|\boldsymbol{x} - \boldsymbol{x}'\|_2. \tag{36}$$

Contradict with the assumption. Thus, the proposed bound can be verified.

$$P(\boldsymbol{x}) = \min_{\boldsymbol{x}'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \geq \frac{\sum_{m=1}^{2} \left(a_y^{(m)} \tilde{W}_{y.}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - a_j^{(m)} \tilde{W}_{j.}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) \right) + \tilde{\beta}_j}{\sqrt{\sum_{m=1}^{2} \left(a_y^{(m)} \tilde{\tau}_y^{(m)} + a_j^{(m)} \tilde{\tau}_j^{(m)} \right)^2}}$$
$$s.t.\ \exists j \neq y,\ \tilde{h}_y(\boldsymbol{x}') = \tilde{h}_j(\boldsymbol{x}'). \tag{37}$$

□

7.3 Smoothing process

In this section, we provide the smoothing process of the margin regularization. Since the imbalance problem is introduced by modality preference, the unpreferred modality can barely obtain discriminative representation. Hence, we aim to enhance the unpreferred uni-modal representation learning, through enhancing its margin. In detail, we focus on the following objectives:

$$\max_{\tilde{W}^{(m)}, \phi^{(m)}} \min_{m; k \neq y} \quad \tilde{W}_{y.}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - \tilde{W}_{k.}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}). \tag{38}$$

This objective aims to improve the learning of the weaker uni-modality by encouraging the correct class and punishing others. For better optimization, we first the smooth minimization among class j through LogSumExp function (Nielsen & Sun, 2016):

$$\max_{\tilde{W}^{(m)}, \phi^{(m)}} \min_{m} \quad \tilde{W}_{y.}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}) - \log \left(\sum_{k \neq y} \exp(\tilde{W}_{k.}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)})) \right), \tag{39}$$

and further, smooth the minimization among modality m, and transform it into minimization problem:

$$\min_{\tilde{W}^{(m)}, \phi^{(m)}} \quad -\log \left(\sum_{m=1}^{2} \frac{\exp(\tilde{W}_{y.}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}))}{\sum_{k \neq y} \exp(\tilde{W}_{k.}^{(m)} \phi^{(m)}(\boldsymbol{x}^{(m)}))} \right). \tag{40}$$

Above all, we extend the regularization to the average of all samples and obtain the objective L_1:

$$\min_{\tilde{W}^{(m)}, \phi^{(m)}} \frac{1}{N} \sum_{i=1}^{N} \log \left(\sum_{m=1}^{2} \frac{\sum_{k \neq y} \exp(\tilde{W}_{k.}^{(m)} \phi^{(m)}(\boldsymbol{x}_i^{(m)}))}{\exp(\tilde{W}_{y.}^{(m)} \phi^{(m)}(\boldsymbol{x}_i^{(m)}))} \right), \tag{41}$$

which explicitly enlarges the uni-modal representation learning to solve the imbalance problem without changing the fusion factor $\boldsymbol{a}^{(m)}$, which can lead to a better uni-modal representation.

7.4 Extensive analysis

Extended to different fusion strategies. In the main manuscript, our analysis mainly focuses on the representative fusion strategy, late fusion, which is widely used in multi-modal research Huang

et al. (2022); Wang et al. (2020). Meanwhile, our method are adaptable to other fusion mechanisms, where the modality-specific representations could interact earlier in the process. Previously, we defined the margin as $\zeta_j^{(m)}(\boldsymbol{x}^{(m)})$, where j is the nearest class to calculate the margin, and the margin is only determined by uni-modal input $\boldsymbol{x}^{(m)}$. To adapt our method for these scenarios including intermediate fusion, we can redefine the representation margin as $\zeta_j^{(m)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$, rather than the previously, indicating that both modalities' input influence the margin. This modification allows us to extend the framework to measure multi-modal perturbations in a more integrated manner. Additionally, we can adapt the definition of the Lipschitz constant in our theoretical analysis here to measure how the multi-modal perturbation influences the margin:

$$|\zeta_j^{(m)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) - \zeta_j^{(m)}(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)})| \leq \tau_j^{(m1)} \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}'^{(1)}\right\|_2 + \tau_j^{(m2)} \left\|\boldsymbol{x}^{(2)} - \boldsymbol{x}'^{(2)}\right\|_2 \tag{42}$$

where $\tau_j^{(m1)}$ represents the Lipschitz constant of modality m from modality 1. This constant can reflect how the alteration in modality 1 influences the margin in modality m. Then following the proof in the manuscript, we can observe that:

$$\begin{aligned} &c_j^{(1)}|\zeta_j^{(1)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) - \zeta_j^{(1)}(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)})| + c_j^{(2)}|\zeta_j^{(2)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) - \zeta_j^{(2)}(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)})| \\ &\leq (c_j^{(1)}\tau_j^{(11)} + c_j^{(2)}\tau_j^{(21)}) \left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}'^{(1)}\right\|_2 + (c_j^{(1)}\tau_j^{(12)} + c_j^{(2)}\tau_j^{(22)}) \left\|\boldsymbol{x}^{(2)} - \boldsymbol{x}'^{(2)}\right\|_2 . \end{aligned} \tag{43}$$

Thus, we can obtain the perturbation bound in this setting:

$$\begin{aligned} \min_{\boldsymbol{x}'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 &\geq \frac{c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) + \beta_j}{\sqrt{(c_j^{(1)}\tau_j^{(11)} + c_j^{(2)}\tau_j^{(21)})^2 + (c_j^{(1)}\tau_j^{(12)} + c_j^{(2)}\tau_j^{(22)})^2}} \\ where \quad j \neq y \quad s.t. \quad & c_j^{(1)}\zeta_j^{(1)}(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)}) + c_j^{(2)}\zeta_j^{(2)}(\boldsymbol{x}'^{(1)}, \boldsymbol{x}'^{(2)}) + \beta_j = 0. \end{aligned} \tag{44}$$

The idea of the proof is similar to the one in Section 7.1.

Extended to three modalities. In this study, our analysis discusses the universal situation of two modalities, and can also be extended to the scenario with more than two modalities. To consider more modalities, the key is to introduce the margin of these modalities's representation. Suppose we have l different modality, the input of the m-th modality is $\boldsymbol{x}^{(m)}$, define the representation margin $\zeta_j^{(m)}(\boldsymbol{x}^{(m)})$, and the corresponding Lipschitz constant $\tau_j^{(m)}$. Thus, our bound can be extended to the following formulation.

$$\begin{aligned} \min_{\boldsymbol{x}'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 &\geq \frac{\sum_{m=1}^{l} c_j^{(m)}\zeta_j^{(m)}(\boldsymbol{x}^{(m)}) + \beta_j}{\sqrt{\sum_{m=1}^{l}(c_j^{(m)}\tau_j^{(m)})^2}} \\ where \quad j \neq y \quad s.t. \quad & \sum_{m=1}^{l} c_j^{(m)}\zeta_j^{(m)}(\boldsymbol{x}'^{(m)}) + \beta_j = 0. \end{aligned} \tag{45}$$

8 Addition for Experiment

8.1 Details for datasets

We conduct experiments on three datasets Kinetics-Sounds (Arandjelovic & Zisserman, 2017), UCF101 (Soomro et al., 2012), and VGGSound (Chen et al., 2020). Here are the details.

Kinetics-Sounds (KS) (Arandjelovic & Zisserman, 2017) is a dataset containing 31 human action classes selected from Kinetics dataset (Kay et al., 2017) which contains 400 classes of YouTube videos. All videos are manually annotated for human action using Mechanical Turk and cropped to 10 seconds long around the action. The 31 classes were chosen to be potentially manifested visually and aurally, such as playing various instruments. This dataset contains 22k 10-second video clips, and for each video, we select 3 frames (first, middle, and last frame)

UCF101 (Soomro et al., 2012) consists of in-the-wild videos from 101 action classes, it is typically regarded as a multi-modal dataset with RGB and optical flow modalities. This dataset contains 13,320 videos of human actions across 101 classes that have previously been used for video synthesis and prediction. For each video, we select the middle frame of RGB and 10 frames of optimal flow for training and prediction. And our backbone is the pre-trained ResNet 18.

VGGSound(VGGS) (Chen et al., 2020) is a large-scale video dataset that contains 309 classes, covering a wide range of audio events in everyday life. All videos in VGGSound are captured "in the wild" with audio-visual correspondence in the sense that the sound source is visually evident. The duration of each video is 10 seconds. In our experimental settings, 168,618 videos are used for training and validation, and 13,954 videos are used for testing because some videos are not available now on YouTube. For each video, we also select the middle frame for prediction.

8.2 Effectiveness Validation

8.2.1 Validation across various models

CRMT with different backbones To evaluate the adaptability and effectiveness of our method across diverse frameworks, we conduct comprehensive experiments on the Kinetic-Sounds dataset utilizing different backbone architectures. These architectures encompass the use of ResNet34 for both Vision (V) and Audio (A) modalities, as well as an integrated approach combining ResNet18 for Vision (V) and a Transformer model for Audio (A). The results in Table 4 of these experiments demonstrate both the improvement and flexibility of our method when implemented across different backbones.

Table 4: Comparative analysis of robustness across different backbones, comparing our method (CRMT-JT) with the joint training (JT) approach.

Backbone	Method	Clean	Missing #v	Missing #a	FGM	PGD-ℓ_2
ResNet34 (V) + ResNet34 (A)	JT	0.6424	0.4528	0.2471	0.3132	0.2863
	CRMT-JT	0.7435	0.5269	0.5705	0.4978	0.4746
Resnet18(V) +Transformer(A)	JT	0.5538	0.3387	0.2703	0.2456	0.2100
	CRMT-JT	0.5807	0.3721	0.3539	0.3067	0.2711

CRMT with intermediate fusion strategies. In this manuscript, our analysis mainly focuses on the widely used late fusion method. To further validate our effectiveness, we delve into a widely used intermediate fusion strategy, the Multi-Modal Transfer Module (MMTM) (Vaezi Joze et al., 2020), in which the interaction occurs among representations at mid-level. We compare MMTM and the one that implements our method (CRMT-MMTM) on the Kinetic-Sounds dataset. The experimental results in Table 5 indicate that our method can also boost the performance and robustness for more multi-modal fusion strategies.

Table 5: Experiment on the intermediate fusion strategy MMTM and with our method.

Method	Clean	Missing #v	Missing #a	FGM	PGD-ℓ_2
MMTM	0.6693	0.4542	0.2028	0.3438	0.3205
CRMT-MMTM	0.6737	0.5211	0.3169	0.3445	0.3358

CRMT on pre-trained transformer. In the primary manuscript, we have verified the performance of our method using transformers trained from scratch. To extend the validation of our approach's effectiveness, we have conducted additional experiments utilizing an ImageNet-pretrained Transformer applied to the Kinetic-Sounds dataset. The findings of this extended analysis, detailed in Table 6, demonstrate that our method retains its effectiveness even when implemented with a pre-trained model. This further verifies the robustness and adaptability of our approach in diverse training scenarios.

Table 6: Experiment on Multi-modal Transformer with pre-trained on KS dataset.

Transformer	Clean	FGM	PGD-ℓ_2
Baseline	0.6788	0.1366	0.0865
CRMT (ours)	0.7406	0.3198	0.2078

8.2.2 VALIDATION UNDER DIFFERENT SETTING

Robustness evaluation under various multi-modal attack. In the Experiment section, we initially describe two multi-modal attack methods: Fast Gradient Method (FGM) and Projected Gradient Descent with ℓ_2 norm (PGD-ℓ_2). To further substantiate the robustness of our approach, this section introduces and evaluates three additional multi-modal attack methods. Firstly, we implement the Multi-modal Embedding Attack (MEA) method as proposed by Zhang et al. (2022), which focuses on shifting the joint representation rather than altering the prediction output. Secondly, we explore the efficacy of our method against the Multi-modal Gaussian Noise (MGN) and the Multi-modal Pixel Missing (MPM) attacks. The results in Table 7 of these expanded experiments demonstrate that our methodology is not only robust to the previously examined attacks (FGM and PGD-ℓ_2) but also capable of defending various multi-modal attacks.

Table 7: Comparison of robustness against different multi-modal attacks on KS dataset.

	JT	GB	OGM	PMR	AT	Mixup	MSEFM	CMRT-JT	CRMT-Mix	CRMT-AT
MGN	0.525	0.462	0.527	0.480	0.524	0.501	0.477	0.560	0.634	0.604
MPM	0.315	0.344	0.446	0.328	0.370	0.467	0.298	0.507	0.570	0.548
MEA	0.340	0.412	0.363	0.455	0.588	0.478	0.500	0.556	0.565	0.689

Extended to the UCF dataset with three modalities. During our analysis and experiments, we consider the universal situation containing two modalities. Our method can also be extended to the scenario with more than two modalities. We conduct experiments on the UCF101 dataset using three modalities: RGB, Optical Flow, and RGB Frame Difference. We compared the performance of methods for both trained from scratch and with pre-training ImageNet-pretrained ResNet18. The experimental results detailed in Table 8. The outcomes demonstrate the effectiveness of our method in enhancing multi-modal robustness in more than two modalities.

Table 8: Comparison on UCF101 with three modalities.

	Training from scratch		Pretrained	
Attack	JT	CRMT-JT	JT	CRMT-JT
Clean	0.4490	0.5640	0.8312	0.8506
FGM	0.4005	0.4567	0.2471	0.4138
PGD-ℓ_2	0.3963	0.4312	0.0783	0.2623

Experiment on Image-Text dataset. We apply experiments on vision-text classification tasks to verify our effectiveness. We utilize a widely used Food101 dataset, which inputs an image-text pair for classification. We employ a Vision Transformer (ViT) as our image encoder and BERT as our text encoder, subsequently concatenating their outputs to achieve a unified joint representation. To evaluate robustness, we apply multiple attacks including modality missing and descent-based attacks (FGM and PGD-ℓ_2). It is important to note that attacks like the Fast Gradient Method (FGM) and Projected Gradient Descent with ℓ_2 norm (PGD-ℓ_2) are typically applied to continuous data. Given that text represents discontinuous data, we focus on implementing these attack methods on the image modality. Our results reveal that the text modality is more critical than the image modality, as its absence significantly impacts model performance, as shown in Table 9. Concentrating on the ℓ_2-norm, we achieve enhanced robustness under both FGM and PGD-ℓ_2 attack. Our method demonstrates a notable performance increase in scenarios where text is absent, though there is a slight decline

in performance when the image modality is missing. This is attributed to the huge performance difference between text and image.

Table 9: Comparison of our proposed CRMT method on Image-Text dataset Food101.

Methods	Clean	FGM on Image	PGD-ℓ_2 on Image	Missing Text	Missing Image
JT	0.8218	0.7603	0.7281	0.0497	0.7831
CRMT-JT	0.8257	0.7656	0.7313	0.0759	0.7779

Comparison with additional robust methods In addition to the comparative analyses presented in the Experiment section, we further extend the validation of our method by contrasting it with additional multi-modal robustness approaches. Specifically, we compare our approach with two recent methods: Robust Multi-Task Learning (RMTL) as described by Ma et al. (2022) and Uni-Modal Ensemble with Missing Modality Adaptation (UME-MMA) proposed by Li et al. (2023). For the RMTL implementation, we adopt a multi-task learning framework to develop a robust multi-modal model, incorporating both full-modal and modality-specific tasks. These methods are applied and compared to the Kinetic-Sounds dataset. As illustrated in Table 10, our method demonstrates superior performance in scenarios involving both modality absence and targeted attacks. These results present the robustness and effectiveness of our approach and the validity of our proposed analysis.

Table 10: Comparison with recent multi-modal attack approaches.

Method	Clean	Missing #v	Missing #a	FGM	PGD-ℓ_2
UME-MMA	0.6999	0.5334	0.4666	0.3394	0.3125
RMTL	0.6672	0.5015	0.2994	0.3641	0.3445
CRMT-JT	0.7580	0.5596	0.5908	0.4906	0.4680

Table 11: Performance against different sizes of FGM attack on UCF101 dataset to evaluate the robustness of different iterations in CRMT.

Attack Size	0.0	0.2	0.4	0.7	1.0	1.5	2.0	2.5	3.0
Iter 1	0.7888	0.5511	0.4605	0.4120	0.3851	0.3841	0.3732	0.3647	0.3562
Iter 2	0.7888	0.5514	0.4589	0.4029	0.3864	0.3851	0.3737	0.3648	0.3559
Iter 4	0.7891	0.5543	0.4660	0.4113	0.3921	0.3881	0.3769	0.3705	0.3629

8.3 Iterations in training procedure

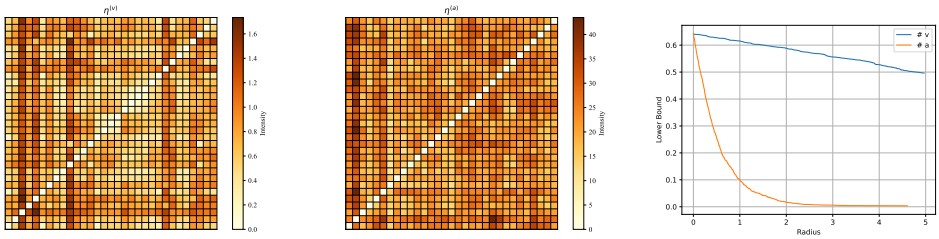
In the method section's concluding part, we currently perform two sequential steps at once. However, it is worth noting that in various applications, incorporating more iterations in the training procedure is a common practice, which holds the potential to further enhance the robustness of our method. To validate the validity of our approach, we conducted experiments by extending our method to incorporate more iterations. As depicted in Table 11, the results of employing additional iterations demonstrate improved robustness capabilities. But only one iteration of our method can achieve considerable robustness, without a huge consumption of training cost.

8.4 Experiments about modality preference

8.4.1 Verification of vulnerable modality

As shown in Figure 3 in the manuscript, we demonstrate how the ratio of the vulnerability indicator (η) varies in our method. Furthermore, to clearly explain this phenomenon, in Figure 8 (a, b), we provide the heat map of the indicator η to represent the robustness of each uni-modality, where the smaller indicator means the modality is more robust. Meanwhile, we also provide the uni-modal perturbation radius to further verify this modality preference, where we list the percentage of safe samples that can always resist this size of perturbation, as shown in Figure 8 (c). It can be seen that

when joint training is applied to the Kinetic-Sounds dataset, the audio modality is definitely more vulnerable than the vision modality, thus explaining the phenomenon in Figure 1 in the paper.

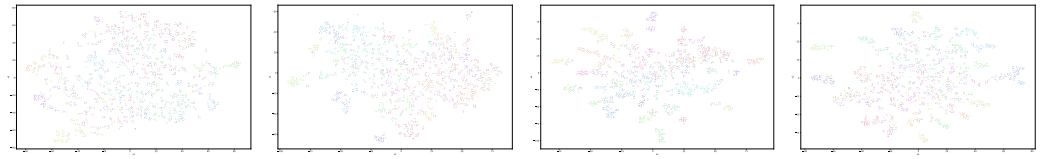


(a) $\eta^{(v)}$ in JT. (b) $\eta^{(a)}$ in JT. (c) Uni-modal perturbation bound.

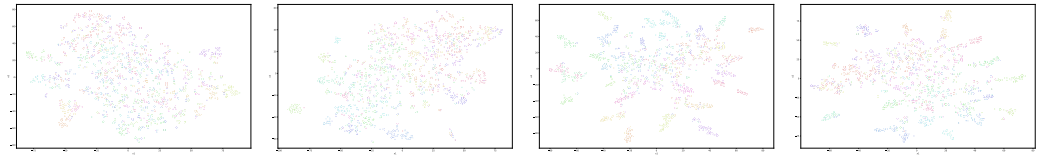
Figure 6: Evaluation of the vulnerability indicators η between modality #v and #a (preferred) and the comparison of safe samples against uni-modal perturbation in the Joint Training (JT) method.

8.4.2 UNI-MODAL REPRESENTATION ANALYSIS

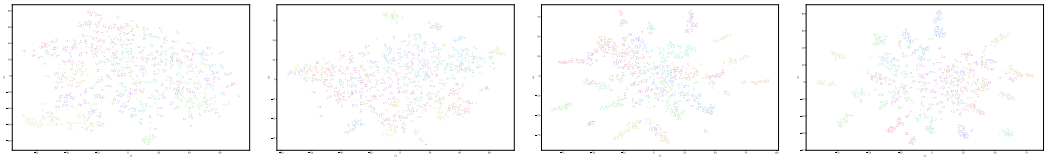
In this part, we verify the imbalanced problem brought by modality preference, which results in the low quality of uni-modal representation. Here, we apply *t-distributed Stochastic Neighbor Embedding* (t-SNE) method on each uni-modal representation and illustrate the visualization of uni-modal representation on KS dataset. As shown in Figure 7, in JT, MMAT, and Mixup, the quality of representation in modality #a is better than modality #v, which implies the margins of representation #v are relatively small. Moreover, by applying our training procedure (see CRMT-JT, CRMT-AT, CRMT-Mix), representations of both modalities show good discrimination, indicating that we enlarge the uni-modal representation margins and eventually enhance the robustness.



(a) JT modality #v. (b) JT modality #a. (c) CRMT-JT modality #v. (d) CRMT-JT modality #a.



(e) MMAT modality #v. (f) MMAT modality #a. (g) CRMT-AT modality #v. (h) CRMT-AT modality #a.

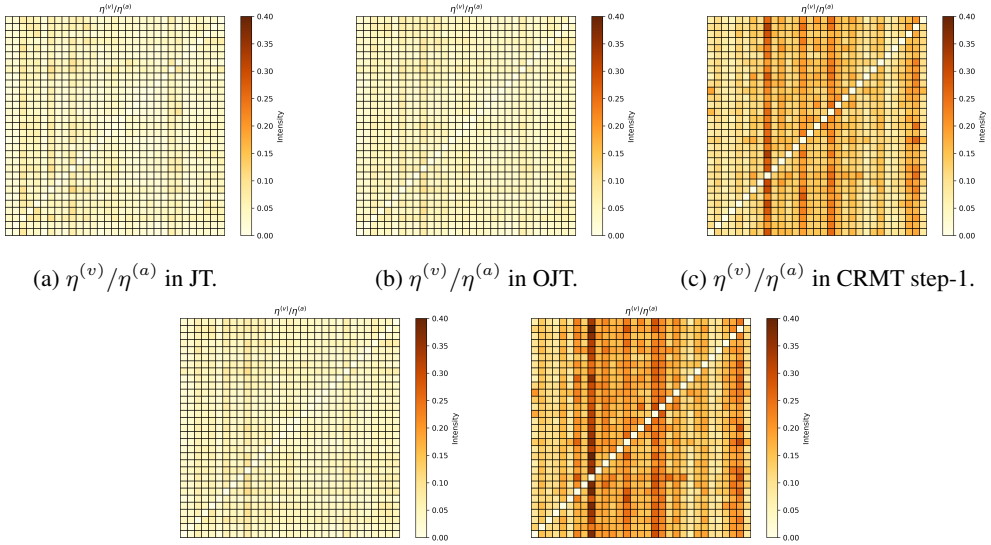


(i) Mixup modality #v. (j) Mixup modality #a. (k) CRMT-Mix modality #v. (l) CRMT-Mix modality #a.

Figure 7: An experiment analyze the uni-modal representation of modality #v and #a (preferred) using t-SNE. We compare different methods and our corresponding CRMT procedure (including CRMT-JT, CRMT-AT, CRMT-Mix) and illustrate that our methods well learn the uni-modal representation.

8.4.3 COMPLEMENT FOR ABLATION STUDIES

As shown in Figure 5, we verify the effectiveness of each step of our training process. Here we provide more evidence that the improved multi-modal robustness is because we alleviate the influence caused by modality preference. Similar to the ablation studies, we report our findings related to solely



(a) $\eta^{(v)}/\eta^{(a)}$ in JT. (b) $\eta^{(v)}/\eta^{(a)}$ in OJT. (c) $\eta^{(v)}/\eta^{(a)}$ in CRMT step-1.

(d) $\eta^{(v)}/\eta^{(a)}$ in CRMT step-2. (e) $\eta^{(v)}/\eta^{(a)}$ in CRMT-JT.

Figure 8: Experiment demonstrates the regularization of the vulnerability indicator in ablation studies.

using the orthogonal framework (OJT), our CRMT focusing exclusively on each step (CRMT-step-1 and CRMT-step-2), and our full method, CRMT-JT, which incorporates both steps. We also apply the ratio of uni-modal robustness indicators, to show the difference of resistance against uni-modal attack. From these heat maps of the ratio in Figure 8, we have the following observations. First, since JT and OJT are different frameworks, their imbalance problem occurs in different class pairs (*i.e.* entries in the heat map). And just applying the orthogonal framework could hardly solve the imbalance problem on the robustness indicator. Second, since the uni-modal representation may not be well learned, just applying step-2 of CRMT only brings marginal improvement. Third, CRMT step-1 can also alleviate the imbalance problem, since it enhances the learning of modality with less discriminative ability, and makes different modality representations more balanced. Fourth, equipped with both enlarging uni-modal representation margin and adjusting integration factors, our method can progressively alleviate the imbalanced problem in attack with different modalities, which further contributes to better robustness.