Vulnerability-Aware Robust Multimodal Adversarial Training

Junrui Zhang[1], **Xinyu Zhao**[2], **Jie Peng**[1], **Chenjie Wang**[3], **Jianmin Ji**[1,*], **Tianlong Chen**[2]

[1]University of Science & Technology of China,
[2]University of North Carolina at Chapel Hill
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{zhangjunrui, pengjieb}@mail.ustc.edu.cn, wangchenjie@iai.ustc.edu.cn,
{xinyu,tianlong}@cs.unc.edu, jianmin@ustc.edu.cn

Abstract

Multimodal learning has shown significant superiority on various tasks by integrating multiple modalities. However, the interdependencies among modalities increase the susceptibility of multimodal models to adversarial attacks. Existing methods mainly focus on attacks on specific modalities or indiscriminately attack all modalities. In this paper, we find that these approaches ignore the differences between modalities in their contribution to final robustness, resulting in suboptimal robustness performance. To bridge this gap, we introduce **V**ulnerability-**A**ware **R**obust **M**ultimodal **A**dversarial **T**raining (VARMAT), a probe-in-training adversarial training method that improves multimodal robustness by identifying the vulnerability of each modality. To be specific, VARMAT first explicitly quantifies the vulnerability of each modality, grounded in a first-order approximation of the attack objective (Probe). Then, we propose a targeted regularization term that penalizes modalities with high vulnerability, guiding robust learning while maintaining task accuracy (Training). We demonstrate the enhanced robustness of our method across multiple multimodal datasets involving diverse modalities. Finally, we achieve $\{12.73\%, 22.21\%, 11.19\%\}$ robustness improvement on three multimodal datasets, revealing a significant blind spot in multimodal adversarial training.

Code — https://github.com/AlniyatRui/VARMAT

1 Introduction

With the rapid advancement of artificial intelligence, multimodal learning (Liang et al. 2023; Yu, Yoon, and Bansal 2025; Lei et al. 2024; Zhang et al. 2025a) has demonstrated powerful perception and reasoning capabilities across a wide range of real-world applications. These abilities are particularly valuable in complex and dynamic environments, where relying on a single modality often fails to provide sufficient information for robust and reliable prediction. For instance, in sentiment analysis, the same sentence can express different emotions depending on how it is spoken, where multimodal features are necessary to provide a comprehensive understanding of the speaker's intent (Zadeh et al. 2018).

Nevertheless, the advantages of multimodal learning are accompanied by significant security challenges. It is widely recognized that deep neural networks are vulnerable to adversarial examples crafted by adding human-imperceptible perturbations, which mislead models into making incorrect predictions (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018). While this vulnerability is inherent in most deep learning models, recent studies indicate that incorporating multiple heterogeneous modalities substantially increases the attack surface (Bagdasaryan et al. 2024; Qi et al. 2024), providing adversaries more opportunities for exploitation. For example, adversarial illusions (Bagdasaryan et al. 2024) have been proposed to construct perturbations that make a target modality's embedding close to an arbitrary, adversary-chosen input in another modality, resulting in conflicts between modalities during task execution.

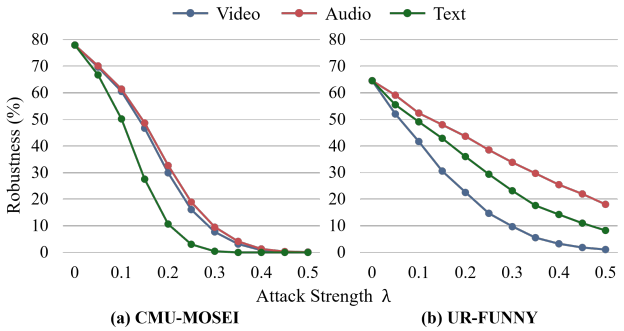

Figure 1: Adversarial robustness under varying attack strengths $\lambda \in [0, 0.5]$ for different modalities, showing significant differences in modality-specific vulnerabilities.

Therefore, it is necessary to design a multimodal defense strategy rather than a single-modal defense to protect against potential adversarial attack threats from arbitrary modalities. Researchers have explored methods to improve the robustness of multimodal models (Luo et al. 2023; Li et al. 2024; Zhang et al. 2025c), among which adversarial training has emerged as one of the most popular defense strategies due to its effectiveness and practicality (Zhang et al. 2019; Pang et al. 2021). Although existing methods have achieved promising results through elaborate attack strategies and adversarial training frameworks, we observe that most of them attack each modality indiscriminately, neglecting the inherent heterogeneity of different modalities (Yin et al. 2023;

*Corresponding Author

Gao et al. 2024; Zheng et al. 2024; Zhang et al. 2025d). Moreover, the design of input-level defenses typically relies on well-defined constraints and standardized input formats, which are often unavailable for some emerging modalities, like the tabular modality in the MIMIC dataset (Johnson et al. 2016), or for modalities that use pre-extracted features, such as point clouds or other 3D structures commonly used to accelerate training (Yu, Yoon, and Bansal 2025).

To address these issues, we explore multimodal robustness through feature-space perturbation. This approach offers a unified view, allowing us to focus on the robustness of each modality without being concerned about the heterogeneities between them. Specifically, given the latent feature of all modalities, we apply consistent constraints and attack algorithms to these features to evaluate the vulnerabilities of each modality individually. As shown in Figure 1, we conduct single-modality PGD attacks (Madry et al. 2018) on the HighMMT model (Liang et al. 2023). We use the CMU-MOSEI (Zadeh et al. 2018) and UR-FUNNY (Hasan et al. 2019) datasets, perturbing one modality at a time to assess the model's robustness. The results show that the model exhibits varying levels of robustness when facing identical adversarial attacks across different modalities in the feature space, indicating that some modalities are more vulnerable than others. This suggests that adversarial training strategies should consider the heterogeneities between modalities, particularly those modalities that are more susceptible to attacks, to achieve better robustness.

Therefore, we first propose a vulnerability-aware attack method to demonstrate that leveraging this discrepancy can craft stronger adversarial examples. Our method reallocates perturbation budgets based on estimated vulnerability, measured by feature magnitude and gradient norm, grounded in a first-order approximation of the attack objective. The perturbation budget refers to the constraint on the maximum allowable perturbation applied to the feature space. By concentrating perturbations on more susceptible modalities, the attack achieves stronger performance, while also revealing the importance of modality-specific vulnerabilities for improving adversarial robustness. Building on this insight, we further propose VARMAT, a vulnerability-aware robust multimodal adversarial training method that directly optimizes the gradient norms of all modalities to dynamically balance and reduce their vulnerabilities, thereby enhancing robustness against adversarial attacks. To be specific, we do not use feature magnitude during training, as directly regularizing it may compromise the expressive capacity of each modality and lead to overfitting under certain attack constraints.

Finally, we evaluate our method on three multimodal datasets from the MultiBench Benchmark (Liang et al. 2021) using the HighMMT (Liang et al. 2023) under various adversarial training settings. In particular, we adopt fast adversarial training methods throughout our experiments, due to their favorable trade-off between robustness and computational efficiency (Shafahi et al. 2019; Wong, Rice, and Kolter 2020; Wang et al. 2024b; Jia et al. 2022). The results demonstrate the effectiveness and efficiency of VARMAT. Specifically, VARMAT achieves robustness improvements of $\{12.73\%, 22.21\%, 11.19\%\}$ on the three datasets, respectively, revealing a critical blind spot in current multimodal adversarial training and underscoring the importance of vulnerability-aware, differentiated optimization to improve robustness. The primary contributions of our research are outlined as follows:

- We systematically reveal the inherent discrepancies in adversarial robustness across different modalities and approximate the attack objective to identify the critical factors that contribute to each modality's vulnerability.
- Building on the identified critical factors, we design a vulnerability-aware attack strategy that reallocates the perturbation budget to concentrate on more vulnerable modalities, achieving better attack performance and underscoring the importance of modality-specific vulnerabilities for improving adversarial robustness.
- We propose VARMAT, a novel vulnerability-aware robust multimodal training method that addresses the previous neglect of inherent discrepancies across modalities. VARMAT incorporates a targeted regularization term that adaptively balances and suppresses the vulnerability of each modality, resulting in better robustness while maintaining the model's expressive capacity.
- Extensive experiments on multiple multimodal datasets demonstrate the effectiveness and efficiency of VARMAT, revealing a significant blind spot in multimodal adversarial training.

2 Related Work

2.1 Multimodal Adversarial Attack

Adversarial attacks refer to deliberately crafted inputs designed to mislead machine learning models into making incorrect predictions. Among various attack strategies, gradient-based methods have emerged as some of the most straightforward and effective approaches, primarily due to their direct exploitation of model vulnerabilities via gradients. FGSM (Goodfellow, Shlens, and Szegedy 2015) pioneered this line of research by demonstrating that even a single-step perturbation along the gradient direction could significantly degrade model performance. Building upon this, iterative variants such as I-FGSM (Kurakin, Goodfellow, and Bengio 2018) were developed to generate more potent adversarial examples through fine-grained, multi-step optimization. To address the challenge of poor local optima, MI-FGSM (Dong et al. 2018) introduced momentum to stabilize gradient updates across iterations, thereby improving attack transferability. Further improvements were achieved by GI-FGSM (Wang et al. 2024a), which incorporates a pre-attack phase to estimate a more informed initial perturbation direction, underscoring the critical role of initialization in crafting effective adversarial examples.

In the context of multimodal learning, multimodal attack strategies exploit interactions between modalities to craft stronger adversarial examples. Co-Attack (Zhang, Yi, and Sang 2022) establishes a framework for synchronized vision-language perturbations through gradient alignment across modalities, demonstrating superior attack success rates compared to unimodal baselines. VLAttack (Yin et al.

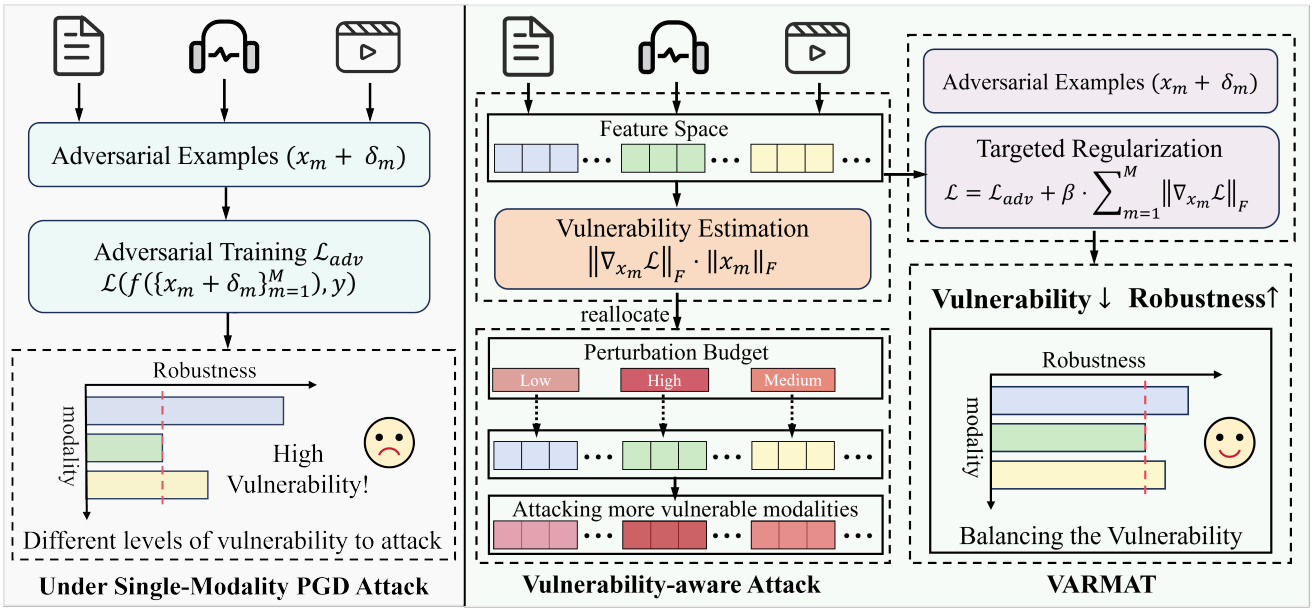

Figure 2: Comparison of vulnerability-aware attack and VARMAT with previous indiscriminate adversarial training.

2023) advances this paradigm by introducing a blockwise similarity attack strategy to generate image perturbations that disrupt universal representations. SGA (Lu et al. 2023) incorporates data augmentation during iterative attack generation, significantly improving cross-model transferability through diversified gradient estimation. VLPTransferAttack (Gao et al. 2025) further enhances transferability by analyzing the intersections of adversarial trajectories between clean and perturbed inputs, thereby preventing overfitting to a specific threat model. The most recent approach, Anyattack (Zhang et al. 2025b), overcomes the scalability limitations of previous methods by allowing any image to serve as a target for attack without requiring label supervision, achieving better performance. While these methods demonstrate effectiveness in vision-language settings, several fundamental limitations persist. First, existing works predominantly focus on image-text pairs, neglecting emerging multimodal systems incorporating audio, video, and other modalities. Second, they largely overlook the strategic question of where to apply perturbations for maximum impact. Current strategies often attack a single predefined modality or indiscriminately perturb all modalities, implicitly assuming that each modality contributes equally to the model's decision and its adversarial vulnerability. In this work, we reveal the inherent differences in vulnerability across modalities, highlighting their differential impact on adversarial robustness.

2.2 Fast Adversarial Training

The most popular defense strategy is PGD-based Adversarial Training (Madry et al. 2018), due to its effectiveness and practicality, where adversarial examples generated using multi-step attacks are used to train a robust model. However, the prohibitive cost of standard PGD-based Adversarial Training motivated a search for more efficient alternatives. This led to the development of Fast Adversarial Training (Wong, Rice, and Kolter 2020), a class of methods aiming to achieve robust models at a computational cost comparable to standard training using FGSM as attack methods. FGSM-RS (Wong, Rice, and Kolter 2020) attempts to mitigate catastrophic overfitting by using a simple initialization with random perturbations. FGSM-GA (Andriushchenko and Flammarion 2020) further improves fast adversarial training by enhancing gradient alignment between clean samples and adversarial examples. FGSM-PCI (Jia et al. 2022) proposes that prior knowledge of adversarial perturbations can effectively guide subsequent adversarial training by using the previous epoch's perturbation as the initialization for further epochs, achieving stable and effective optimization directions. The most recent in this line of research, FGSM-PCO (Wang et al. 2024b), introduces a tailored loss function within the training framework to facilitate the recovery of an overfitted model for effective training, thereby preventing the collapse of the inner optimization loop. While Fast Adversarial Training has matured significantly in the context of unimodal models, particularly in computer vision, the principles and techniques are now being considered for the next frontier of AI: multimodal models.

3 Methodology

In this section, we first introduce the setup of our work in Section 3.1. Then we identify the critical factors contributing to the differences in vulnerability across modalities in Section 3.2. Futhermore, we propose a vulnerability-aware attack method to highlight the impact of varying vulnerabilities on robustness and present VARMAT to achieve a stronger defense in Section 3.3.

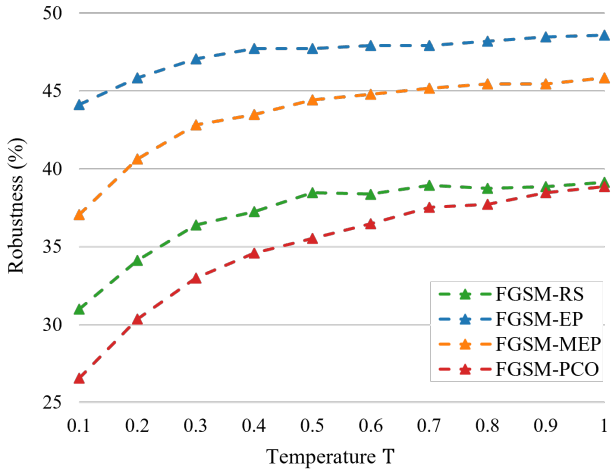

Figure 3: Comparison of robustness between different methods in different temperatures.

3.1 Preliminary

Formally, let $f_\theta(\mathrm{x}) \to y$ denote a multimodal model, where θ represents the model parameters, $\mathrm{x} = \{x_1, x_2, \ldots, x_M\}$ comprises inputs from M modalities, and y is the ground-truth label. Adversarial training is typically formulated as a Min-Max optimization problem:

$$\arg\min_{\theta} \mathbb{E}_{(\mathrm{x},y)} \arg\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(f_\theta(\mathrm{x}+\delta), y), \tag{1}$$

where the inner maximization aims to generate imperceptible perturbations $\delta = \{\delta_1, \delta_2, \ldots, \delta_M\}$ that mislead the model's prediction, and the outer minimization leverages these adversarial examples as training samples to improve the model's robustness against adversarial attacks. The constraint $\|\delta\|_p \leq \epsilon$ would project $\mathrm{x}+\delta$ to the ℓ_p ball around x with radius ϵ. In this paper, we construct perturbations in the feature space from a unified perspective, where the input x represents encoded features across different modalities.

To thoroughly investigate the varying vulnerabilities across different modalities, we simultaneously study multimodal adversarial attacks and defenses. We start by analyzing the approximation attack objective of inner maximization of multimodal fast adversarial training in Section 3.2. Motivated by this analysis, we reallocate the perturbation budget across modalities based on modality vulnerability weights w_m to construct the vulnerability-aware adversarial attack. The overall constraint becomes:

$$\|\delta_m\|_p \leq w_m \cdot \epsilon_m, \quad \text{s.t.} \sum_{m=1}^{M} w_m = 1, \tag{2}$$

However, in fast adversarial training, recent approaches often leverage prior information to prevent catastrophic overfitting. Therefore, directly allocating the perturbation budget across epochs may lead to unstable and suboptimal training. To address this issue, we introduce a regularization term based on the theoretical approximation attack rather than directly allocating the perturbation budget. Detailed attack and defense strategies are described in Section 3.3.

3.2 Approximation for $Frobenius$-norm Attacks

Following prior work (Zhu et al. 2020), we constrain multimodal feature-space perturbations by the $Frobenius$-norm:

$$\epsilon_m = \lambda \cdot \|x_m\|_F, \tag{3}$$

where λ represents the strength of the attack. We then consider a first-order approximation of the inner maximization objective in Eq. (1).

$$\mathcal{L}(f_\theta(\mathrm{x}+\delta), y) \approx \mathcal{L}(f_\theta(\mathrm{x}), y) + \delta \cdot \nabla_{\mathrm{x}} \mathcal{L}(f_\theta(\mathrm{x}), y), \tag{4}$$

For simplicity, we denote $\mathcal{L}(f_\theta(\mathrm{x}), y)$ as $\mathcal{L}$. The gradient of a natural sample is used in fast adversarial training to identify the most vulnerable direction for generating single-step attacks. Thus, the first-order term $\delta \cdot \nabla_{\mathrm{x}} \mathcal{L}(f_\theta(\mathrm{x}), y)$ in Eq. (4) approximates the loss increase, denoted as $\Delta\mathcal{L}$. Since feature are not explicitly constrained by upper or lower bounds, under the $Frobenius$-norm constraint, the single-step perturbation of modality m can be formulated as:

$$\delta_m = \frac{\nabla_{x_m} \mathcal{L}}{\|\nabla_{x_m} \mathcal{L}\|_F} \cdot \epsilon_m \cdot w_m, \tag{5}$$

Consequently, the approximate loss increase $\Delta\mathcal{L}$ can be expressed as:

$$\Delta\mathcal{L} = \sum_{m=1}^{M} (\|\nabla_{x_m} \mathcal{L}\|_F \cdot \|x_m\|_F \cdot \lambda \cdot w_m), \tag{6}$$

We observe that modality vulnerability is jointly influenced by both input features and their corresponding gradients. As illustrated in Figure 2, previous methods typically adopt indiscriminate strategies that neglect inherent differences among modalities. To highlight this issue, we design a vulnerability-aware adversarial attack that reallocates the perturbation budget based on modality-specific vulnerabilities, resulting in stronger attacks. Motivated by this insight, we further propose `VARMAT`, which employs targeted vulnerability-aware regularization to balance vulnerabilities across modalities during adversarial training, thereby enhancing overall robustness while preserving the model's expressive capacity.

3.3 Vulnerability-Aware Adversarial Rosbustness

According to Eq. (6), we approximate each modality's vulnerability using the $\|\nabla_{x_m} \mathcal{L}\|_F \cdot \|x_m\|_F$. This estimation can be efficiently computed, as both x_m and $\nabla_{x_m} \mathcal{L}$ can be obtained through a single forward and backward pass, providing an efficient and lightweight way to probe the relative vulnerability across modalities. Moreover, the probing process is independent of the attack method, which ensures the generality and flexibility of `VARMAT`. Consequently, based on this estimation, we compute the modality-specific vulnerability weights w_m using the following normalization:

$$w_m = \frac{\|\nabla_{x_m} \mathcal{L}\|_F \cdot \|x_m\|_F}{\sum_{k=1}^{M} (\|\nabla_{x_k} \mathcal{L}\|_F \cdot \|x_k\|_F)}, \tag{7}$$

To further enhance controllability, we apply a temperature-scaled softmax to the normalized weights:

$$w_m = \text{Softmax}\,(w_m / T) \tag{8}$$

Algorithm 1: Vulnerability-Aware FGSM-RS

Input:Target model f_θ, training epoch $\mathcal{T}$, mini-batch data $\mathcal{B}$, modalties M, learning rate γ, attack strength λ, regularization coefficient β.

Output: Model parameter θ

for $t = 1, \ldots, \mathcal{T}$ **do**
  for $\{\mathrm{x}, y\} \sim \mathcal{B}$ **do**
    for $m = 1, \ldots, M$ **do**
      $\epsilon_m \leftarrow \frac{\lambda}{M} \cdot \|x_m\|_F$
      if $t == 1$ **then**
        $\delta_m \leftarrow \mathcal{U}_{[-\epsilon_m, \epsilon_m]}$
        $\delta_m \leftarrow \delta_m \cdot \min\left(1, \epsilon_m / \|\delta_m\|_F\right)$
      end if
    end for
    $\mathcal{L}_{Reg} \leftarrow 0$
    for $m = 1, \ldots, M$ **do**
      $g_t \leftarrow \nabla_{x_m} \mathcal{L}(f_\theta(\mathrm{x}), y)$
      $g_{adv} \leftarrow \nabla_{x_m} \mathcal{L}(f_\theta(\mathrm{x} + \delta), y)$
      $\delta_m \leftarrow \delta_m + g_{adv} / \|g_{adv}\|_F \cdot \epsilon_m$
      $\delta_m \leftarrow \delta_m \cdot \min\left(1, \epsilon_m / \|\delta_m\|_F\right)$
      $\mathcal{L}_{Reg} \leftarrow \mathcal{L}_{Reg} + \beta \cdot \|g_t\|_F$
    end for
    $\theta_{t+1} \leftarrow \theta_t + \frac{\gamma}{|\mathcal{B}|} \sum^{\mathcal{B}} \nabla(\mathcal{L}(f_\theta(\mathrm{x} + \delta), y) + \mathcal{L}_{Reg})$
  end for
end for
return θ

where T is a temperature hyperparameter that modulates the sharpness of the distribution. As $T \to 0$, the perturbation budget becomes increasingly concentrated on the most vulnerable modality; as $T \to \infty$, the weights approach a uniform distribution. As shown in Figure 3, we adjust the temperature to examine how adversarially trained models perform under different perturbation allocation patterns. The results indicate that indiscriminate training methods struggle to accommodate modality-specific vulnerabilities across heterogeneous modalities.

Ideally, directly applying this attack strategy during adversarial training could help the model improve its robustness on more vulnerable modalities due to the adaptive reallocation strategy. Although such a strategy was applicable in earlier methods, it conflicts with recent fast adversarial training techniques, which typically incorporate prior knowledge across training epochs to refine the attack process and mitigate catastrophic overfitting.

To avoid this conflict, we propose incorporating a regularization term that directly optimizes the approximated attack objective $\sum_{m=1}^{M}(\|\nabla_{x_m}\mathcal{L}\|_F \cdot \|x_m\|_F)$ in Eq. (6). Since different modalities contribute unequally to this term, more vulnerable modalities naturally receive stronger optimization weight. As a result, the model implicitly prioritizes these modalities during training, promoting a better alignment between vulnerability and robustness learning. However, this formulation introduces a potential optimization trap: minimizing the regularization objective may lead to a degenerate solution where $\|x_m\|_F \to 0$. Such collapse impairs the model's ability to maintain expressive representations and also increases the risk of overfitting. To mitigate this issue, we redesign the regularization term to focus solely on the gradient component, thereby avoiding direct penalization of the feature norm. The revised formulation is:

$$\mathcal{L}_{Reg} = \beta \cdot \sum_{m=1}^{M} \|\nabla_{x_m}\mathcal{L}\|_F \tag{9}$$

where the coefficient β controls the strength of regularization. As shown in Algorithm 1, we provide a deployment using FGSM-RS, where the fast adversarial training framework is designed to be easily adapted to other algorithms.

4 Experiments

4.1 Experimental Setup

Datesets and Model Our experiments are conducted on three diverse multimodal datasets from the MultiBench benchmark (Liang et al. 2021): CMU-MOSEI (Zadeh et al. 2018), UR-FUNNY (Hasan et al. 2019), and AVMNIST (Vielzeuf et al. 2018). CMU-MOSEI and UR-FUNNY include video, text, and audio modalities, while AVMNIST consists of image and audio modalities. For all experiments, we utilize HighMMT (Liang et al. 2023) as our backbone.

Compared Methods We integrate our method with current state-of-the-art fast adversarial training approaches to demonstrate its effectiveness, including FGSM-RS (Wong, Rice, and Kolter 2020), FGSM-EP (Jia et al. 2022), FGSM-MEP (Jia et al. 2022), and FGSM-PCO (Wang et al. 2024b). Except for FGSM-RS, other methods incorporate prior information from earlier training epochs.

Training Details For all compared methods, the attack strength is set to $\lambda = 0.01$ during training and $\lambda \in \{0.2, 0.5\}$ during testing. For multi-step attacks, the step size is set to $\alpha = \epsilon/I$, with $I = 10$ iterations. We select the best model checkpoints based on the highest robustness on the validation set under $\lambda = 0.2$, evaluated using the PGD attack. Following the original configurations, we use the Adam optimizer with a learning rate of 0.0008 and a weight decay of 0.01. All models are trained on 1 NVIDIA A30 GPU for 10 epochs. The hyperparameters used in the fast adversarial training methods are consistent with their original settings. The regularization coefficient β in our method is determined based on the gradient magnitude of the undefended model trained on each dataset. Specifically, we set $\beta = 1000$ for CMU-MOSEI, and $\beta = 50$ for both UR-FUNNY and AVMNIST.

Evaluation Metrics We evaluate the robustness of all adversarially trained models under FGSM and PGD attacks. To be specific, following the standard robust evaluation (Croce and Hein 2020), we compute robustness as the ratio of samples that are correctly classified on both the clean and adversarial examples over the entire dataset, in order to avoid overestimating robustness.

Method	Variant	Clean	CMU-MOSEI ($\lambda = 0.2$)				CMU-MOSEI ($\lambda = 0.5$)			
			FGSM	PGD	V-FGSM	V-PGD	FGSM	PGD	V-FGSM	V-PGD
Undefended	-	77.92	27.78	22.87	26.45	21.69	0.11	0.00	0.00	0.00
FGSM-RS	baseline	78.83	72.00	71.53	71.85	71.40	57.72	55.76	56.99	55.12
	VARMAT	79.15	75.51	75.17	74.35	74.03	69.80	68.49	66.57	64.91
FGSM-EP	baseline	79.09	73.75	73.51	73.01	72.50	65.13	63.77	60.78	58.86
	VARMAT	78.07	75.17	74.86	73.81	73.40	70.23	68.58	66.64	65.00
FGSM-MEP	baseline	78.70	73.72	73.21	72.17	71.63	65.17	63.79	60.41	58.65
	VARMAT	79.30	75.64	75.36	74.50	74.05	69.63	67.46	66.40	64.20
FGSM-PCO	baseline	79.26	73.34	72.91	70.84	70.13	63.13	61.73	56.26	53.63
	VARMAT	78.72	74.28	74.03	71.20	70.64	67.28	65.58	59.72	58.15

Table 1: Robustness (%) under different attack settings on CMU-MOSEI dataset. Baseline refers to the model trained with standard adversarial training without incorporating VARMAT.

Method	Variant	UR-FUNNY ($\lambda = 0.5$)					AVMNIST ($\lambda = 0.5$)				
		Clean	FGSM	PGD	V-FGSM	V-PGD	Clean	FGSM	PGD	V-FGSM	V-PGD
Undefended	-	64.56	7.09	5.20	5.48	3.78	69.04	2.57	0.12	0.74	0.00
FGSM-RS	baseline	65.69	42.25	40.45	39.79	38.19	69.38	6.11	1.78	3.23	0.33
	VARMAT	63.80	57.75	56.43	55.86	53.97	63.21	8.36	3.08	4.02	0.80
FGSM-EP	baseline	64.27	50.38	49.43	48.87	47.73	67.72	4.00	0.35	1.55	0.06
	VARMAT	62.76	57.75	57.18	56.05	54.54	62.16	15.16	11.08	12.74	8.53
FGSM-MEP	baseline	63.80	48.20	46.50	46.60	44.52	68.98	7.37	1.22	3.65	0.22
	VARMAT	63.33	55.48	54.25	53.31	52.36	63.40	10.18	4.50	7.79	2.63
FGSM-PCO	baseline	62.48	44.05	41.21	40.55	35.82	68.10	4.77	1.17	1.95	0.37
	VARMAT	63.89	59.83	59.55	58.60	58.03	66.76	12.46	5.70	6.22	0.84

Table 2: Robustness (%) under different attack settings on UR-FUNNY and AVMNIST datasets.

4.2 Comparative Results

Results on Multimodal Datasets To demonstrate the effectiveness of VARMAT, we compare several state-of-the-art fast adversarial training algorithms on the CMU-MOSEI, UR-FUNNY, and AVMNIST datasets. The comparison results on CMU-MOSEI are shown in Table 1, where the baseline refers to the model trained with standard adversarial training without incorporating VARMAT. We report the robustness under different attack settings as well as the accuracy on the clean samples. Additionally, we evaluate the robustness of each adversarial training method against our proposed vulnerability-aware attack, denoted as V-FGSM and V-PGD, with the temperature $T = 0.5$. Overall, VARMAT achieves consistently higher robustness across varying attack strengths and attack methods. The improvements on the four methods reach up to $\{12.73\%, 6.14\%, 5.99\%, 4.52\%\}$, respectively, which demonstrates that VARMAT effectively enhances robustness by quantifying and regularizing the vulnerability of each modality. Moreover, our vulnerability-aware attack leads to a more severe degradation in robustness for both clean and adversarially trained models, highlighting the significance of vulnerability differences and their impact on model robustness. To further validate the generalizability of our method, we conduct additional experiments on the UR-FUNNY and AVMNIST datasets where VARMAT achieves consistent robustness improvements on these datasets. As shown in Table 2, in the UR-FUNNY four methods improvements reach up to $\{16.07\%, 7.75\%, 7.84\%, 22.21\%\}$ and in the AVMNIST four methods improvements reach up to $\{2.25\%, 11.19\%, 4.14\%, 7.69\%\}$.

Single-Modality Attack To further demonstrate the consistent robustness of VARMAT across modalities, we conduct a single-modality attack experiment on CMU-MOSEI using the PGD attack. As shown in Figure 5, VARMAT significantly improves robustness for high-vulnerability modalities while maintaining performance in others. This demonstrates that VARMAT effectively strengthens the defense of high-vulnerability modalities , enhancing overall robustness.

Efficiency of VARMAT The most critical advantage of fast adversarial training lies in the efficiency of its single-step attack. To assess the computational overhead introduced by VARMAT, we measure the average training time per batch across different methods on CMU-MOSEI, as shown in Figure 4. The results show that the additional overhead remains constant and negligible across various methods. Meanwhile, as attack method becomes more complex or adopts multi-step attacks, the relative time cost introduced by VARMAT

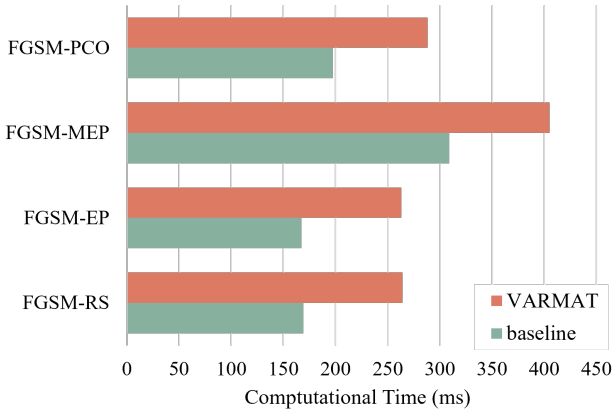

Figure 4: Comparison of computational time between fast adversarial training methods.

Strategy	FGSM-RS	FGSM-EP	FGSM-MEP	FGSM-PCO
S1	62.63	66.79	65.15	63.45
S2	35.54	17.51	52.14	26.60
S3	51.88	59.96	62.76	64.98
Trap	75.83	74.52	74.31	72.80
VARMAT	68.49	68.58	67.46	65.58

Table 3: Ablation of different adversarial training strategies under PGD attack ($\lambda = 0.5$).

further decreases. This demonstrates that VARMAT improves robustness while maintaining high training efficiency.

4.3 Ablation Study

We conduct an ablation study on CMU-MOSEI dataset with different training strategies to demonstrate the effectiveness of VARMAT:

- S1: Regularize only the most vulnerable modality in the current epoch, rather than applying regularization to all modalities simultaneously.
- S2: Utilize our vulnerability-aware attack to directly generate adversarial examples, where the perturbation budget for each modality may change in every epoch.
- S3: Attack only the most vulnerable modality, which is identified by applying the attack on the undefended model (i.e., Video modality in Figure 1).
- Trap: Optimize the $\|\mathrm{x}\|_F \cdot \|\nabla_{\mathrm{x}}\mathcal{L}\|_F$, which may lead to overfitting and degraded expressive capability.

As shown in Table 3, except for the Trap strategy, VARMAT consistently outperforms other baselines across different methods. This suggests that the proposed regularization effectively enhances overall robustness. However, the Trap strategy maintains robustness under attack that is nearly equivalent to its performance on clean inputs, suggesting possible overfitting to the specific attack constraints.

To further investigate this phenomenon, we conduct an additional experiment in the input space. Specifically, we directly transfer the attack constraints from the feature space

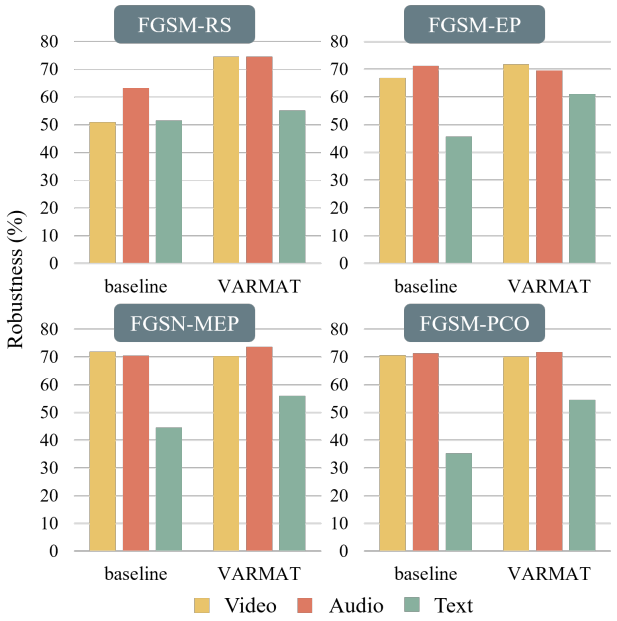

Figure 5: Comparison of single-modality PGD attack performance across fast adversarial training methods ($\lambda = 0.5$).

Strategy	FGSM-RS	FGSM-EP	FGSM-MEP	FGSM-PCO
Trap	7.15	29.51	13.74	2.76
VARMAT	9.58	50.03	39.33	53.67

Table 4: Robustness (%) of Trap and VARMAT under the input-space PGD attack ($\lambda = 0.1$).

to the input space. Although this inevitably generates adversarial examples that may exceed the natural bounds of individual modalities, we argue that the results still provide meaningful insights into the underlying optimization trap. As shown in Table 4, the Trap strategy exhibits severe overfitting across various fast adversarial training methods. This suggests that although feature magnitude may correlate with vulnerability in a frozen model, directly optimizing it leads to undesired overfitting and should be avoided.

5 Conclusion

In this paper, we reveal that inherent differences in modality-specific vulnerabilities have a significant impact on model robustness under adversarial attacks. We demonstrate that leveraging these differences enables the construction of more targeted and effective adversarial attacks. To address this issue, we propose VARMAT, a novel vulnerability-aware robust multimodal adversarial training method that introduces a regularization term to reduce and balance vulnerabilities across modalities. Experimental results on various multimodal datasets validate the effectiveness and efficiency of VARMAT, and further highlight the importance of differentiated training based on the unequal contributions of each modality to improving adversarial robustness.

References

Andriushchenko, M.; and Flammarion, N. 2020. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33: 16048–16059.

Bagdasaryan, E.; Jha, R.; Shmatikov, V.; and Zhang, T. 2024. Adversarial illusions in {Multi-Modal} embeddings. In *33rd USENIX Security Symposium (USENIX Security 24)*, 3009–3025.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Gao, S.; Jia, X.; Ren, X.; Tsang, I.; and Guo, Q. 2024. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *European Conference on Computer Vision*, 442–460. Springer.

Gao, S.; Jia, X.; Ren, X.; Tsang, I.; and Guo, Q. 2025. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *European Conference on Computer Vision*, 442–460. Springer.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Hasan, M. K.; Rahman, W.; Bagher Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; and Hoque, M. E. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2046–2056. Hong Kong, China: Association for Computational Linguistics.

Jia, X.; Zhang, Y.; Wei, X.; Wu, B.; Ma, K.; Wang, J.; and Cao, X. 2022. Prior-guided adversarial initialization for fast adversarial training. In *European Conference on Computer Vision*, 567–584. Springer.

Johnson, A.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database, Sci. *Data*, 3(1): 1–9.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Lei, W.; Ge, Y.; Yi, K.; Zhang, J.; Gao, D.; Sun, D.; Ge, Y.; Shan, Y.; and Shou, M. Z. 2024. Vit-lens: Towards omni-modal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26647–26657.

Li, L.; Guan, H.; Qiu, J.; and Spratling, M. 2024. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24408–24419.

Liang, P. P.; Lyu, Y.; Fan, X.; Tsaw, J.; Liu, Y.; Mo, S.; Yogatama, D.; Morency, L.-P.; and Salakhutdinov, R. 2023. High-Modality Multimodal Transformer: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning. *Transactions on Machine Learning Research*.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M. A.; Zhu, Y.; et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, 2021(DB1): 1.

Lu, D.; Wang, Z.; Wang, T.; Guan, W.; Gao, H.; and Zheng, F. 2023. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 102–111.

Luo, H.; Gu, J.; Liu, F.; and Torr, P. 2023. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Pang, T.; Yang, X.; Dong, Y.; Su, H.; and Zhu, J. 2021. Bag of Tricks for Adversarial Training. In *International Conference on Learning Representations*.

Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 21527–21536.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *Advances in neural information processing systems*, 32.

Vielzeuf, V.; Lechervy, A.; Pateux, S.; and Jurie, F. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Wang, J.; Chen, Z.; Jiang, K.; Yang, D.; Hong, L.; Guo, P.; Guo, H.; and Zhang, W. 2024a. Boosting the transferability of adversarial attacks with global momentum initialization. *Expert Systems with Applications*, 255: 124757.

Wang, Z.; Wang, H.; Tian, C.; and Jin, Y. 2024b. Preventing catastrophic overfitting in fast adversarial training: A bi-level optimization perspective. In *European Conference on Computer Vision*, 144–160. Springer.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*.

Yin, Z.; Ye, M.; Zhang, T.; Du, T.; Zhu, J.; Liu, H.; Chen, J.; Wang, T.; and Ma, F. 2023. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36: 52936–52956.

Yu, S.; Yoon, J.; and Bansal, M. 2025. CREMA: Generalizable and Efficient Video-Language Reasoning via Multimodal Modular Fusion. In *The Thirteenth International Conference on Learning Representations*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, J.; Wang, C.; Peng, J.; Li, H.; Ji, J.; Zhang, Y.; and Zhang, Y. 2025a. CAFE-AD: Cross-Scenario Adaptive Feature Enhancement for Trajectory Planning in Autonomous Driving. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 1300–1307.

Zhang, J.; Ye, J.; Ma, X.; Li, Y.; Yang, Y.; Chen, Y.; Sang, J.; and Yeung, D.-Y. 2025b. AnyAttack: Towards Large-scale Self-supervised Adversarial Attacks on Vision-language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19900–19909.

Zhang, J.; Yi, Q.; and Sang, J. 2022. Towards Adversarial Attack on Vision-Language Pre-training Models. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Zhang, M.; Bi, K.; Chen, W.; Guo, J.; and Cheng, X. 2025c. CLIPure: Purification in Latent Space via CLIP for Adversarially Robust Zero-Shot Classification. In *The Thirteenth International Conference on Learning Representations*.

Zhang, Z.; Liang, S.; Shimada, D.; and Xu, C. 2025d. Rethinking Audio-Visual Adversarial Vulnerability from Temporal and Modality Perspectives. In *The Thirteenth International Conference on Learning Representations*.

Zheng, H.; Deng, X.; Jiang, W.; and Li, W. 2024. A unified understanding of adversarial vulnerability regarding unimodal models and vision-language pre-training models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 18–27.

Zhu, C.; Cheng, Y.; Gan, Z.; Sun, S.; Goldstein, T.; and Liu, J. 2020. FreeLB: Enhanced Adversarial Training for Natural Language Understanding. In *International Conference on Learning Representations*.