# MMCert: Provable Defense against Adversarial Attacks to Multi-modal Models

Yanting Wang[1], Hongye Fu[2*], Wei Zou[1], and Jinyuan Jia[1]
[1]The Pennsylvania State University, [2]Zhejiang University
[1]{yanting, weizou, jinyuan}@psu.edu, [2]3200102866@zju.edu.cn

## Abstract

*Different from a unimodal model whose input is from a single modality, the input (called multi-modal input) of a multi-modal model is from multiple modalities such as image, 3D points, audio, text, etc. Similar to unimodal models, many existing studies show that a multi-modal model is also vulnerable to adversarial perturbation, where an attacker could add small perturbation to all modalities of a multi-modal input such that the multi-modal model makes incorrect predictions for it. Existing certified defenses are mostly designed for unimodal models, which achieve sub-optimal certified robustness guarantees when extended to multi-modal models as shown in our experimental results. In our work, we propose MMCert, the first certified defense against adversarial attacks to a multi-modal model. We derive a lower bound on the performance of our MMCert under arbitrary adversarial attacks with bounded perturbations to both modalities (e.g., in the context of auto-driving, we bound the number of changed pixels in both RGB image and depth image). We evaluate our MMCert using two benchmark datasets: one for the multi-modal road segmentation task and the other for the multi-modal emotion recognition task. Moreover, we compare our MMCert with a state-of-the-art certified defense extended from unimodal models. Our experimental results show that our MMCert outperforms the baseline.*

## 1. Introduction

With the rapid advancement of machine learning, multi-modal models have emerged as a powerful paradigm. Differing from their unimodal counterpart whose input is from a singular modality, these multi-modal models leverage input (called *multi-modal input*) from diverse modalities such as images, 3D data points, audio, and text [8, 11, 24, 39, 50]. Those multi-modal models have been widely used in many security and safety critical applications such as autonomous driving [11, 30, 35, 44, 48] and medical imaging [15].

As shown in many existing studies [13, 37, 42], unimodal models are susceptible to adversarial attacks. There is no exception for multi-modal models. In particular, many recent studies [6, 41, 43, 45, 54, 61] showed that multi-modal models are also vulnerable to adversarial perturbations. In particular, an attacker could simultaneously manipulate all modalities of a multi-modal input such that a multi-modal model makes incorrect predictions. For instance, in the scenario of road segmentation for auto-driving, the attacker can add small perturbations to both the RGB image (captured by a camera) and the depth image (captured by a LiDAR depth sensor) to degrade the segmentation quality. Similarly, in the scenario of video emotion recognition, the attacker can apply subtle disruptions to both visual and audio data to reduce prediction accuracy.

Many defenses were proposed to defend against adversarial attacks, In particular, they can be categorized into *empirical defenses* [25, 34, 40, 45, 49, 52, 56] and *certified defenses* [7, 10, 14, 19, 27, 28, 51, 53, 57, 60]. Many existing studies [4, 5, 46] showed that most empirical defenses could be broken by strong, adaptive attacks (one exception is adversarial training [34]). Therefore, we focus on certified defense in this work. Existing certified defenses are mainly designed for unimodal models (its input is from a single modality). Our experimental results show that they achieve sub-optimal performance when extended to defend against adversarial attacks for multi-modal models. The key reason is that when the attacker adds $l_p$ bounded perturbations to all modalities, the space of perturbed multi-modal inputs cannot be simply formulated as a $l_p$ ball. In this work, we focus on $l_0$-like adversarial attacks applied to each modality (i.e., manipulate a certain number of features for each modality) due to their straightforward applicability across various modalities. The investigation of alternative forms of attacks is reserved for future research.

**Our work.** We propose MMCert, the first certified defense against adversarial attacks to multi-modal models. Suppose we have a multi-modal input $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_T)$ with $T$ modalities, where $\mathbf{m}_i$ contains a set/sequence of basic elements from the $i$-th modality. We consider a general scenario, where each element could be arbitrary. For in-

*Hongye Fu performed this research when he was a remote intern.

stance, each element could be a pixel value, a 3D point, an image frame, an audio frame, etc.. Given a multi-modal input $\mathbf{M}$ and a multi-modal model $g$ (called *base multi-modal model*), we first create multiple sub-sampled multi-modal inputs. In particular, each sub-sampled multi-modal input is obtained by randomly sub-sampling $k_1, k_2, \cdots, k_T$ basic elements from $\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_T$, respectively. Then, we use the base multi-modal model $g$ to make a prediction for each sub-sampled multi-modal input. Finally, we build an ensemble multi-modal model by aggregating those predictions as the final prediction made by our ensemble multi-modal classifier for the given multi-modal input $\mathbf{M}$.

We derive the provable robustness guarantee of our ensemble multi-modal model. In particular, we show that our ensemble multi-modal model provably makes the same prediction for a multi-modal input when the number of added (or deleted or modified) basic elements to $\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_T$ is no larger than $r_1, r_2, \cdots, r_T$. Intuitively, there is a considerable overlap between the space of randomly sub-sampled multi-modal inputs before the attack and those sub-sampled after the attack. This suggests that the alterations in the output prediction probabilities are constrained. Following [10, 20], the robustness guarantee is achieved by utilizing Neyman-Pearson Lemma [36].

We conduct a systematic evaluation for our MMCert on two benchmark datasets for multi-modal road segmentation and multi-modal emotion recognition tasks, respectively. We measure the performance lower bounds of our defense under adversarial attacks, with the constraint that the number of modified (or deleted or added) basic elements to each modality is bounded. We compare our MMCert with randomized ablation [28], which is a state-of-the-art certified defense for unimodal models. Our experimental results show that our MMCert significantly outperforms randomized ablation when extending it to multi-modal models.

In summary, we make the following major contributions:

- We propose MMCert, the *first* certified defense against adversarial attacks to multi-modal models.
- We derive the provable robustness guarantees of our MMCert.
- We conduct a systematic evaluation for our MMCert and compare it with state-of-the-art certified defense for unimodal models.

## 2. Background and Related Work

Multi-modal models [8, 11, 24, 39] are designed to process information across multiple types of data, such as text, images, 3D point clouds, and audio, simultaneously. Multi-modal models have shown impressive results across a variety of applications, such as scene understanding [24], object detection [16, 39, 47], sentiment analysis [8, 26, 58, 59], visual question answering [3, 18], and semantic segmentation [11, 31].

For simplicity, we use $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_T)$ to denote a multi-modal input with $T$ modalities, where $\mathbf{m}_i$ represents the group of basic elements (pixels, images, audio) from the $i$th ($i = 1, 2, \cdots, T$) modality.

### 2.1. Adversarial Attacks to Multi-modal Models

Many existing studies [6, 41, 43, 45, 54, 61] showed that multi-modal models are vulnerable to adversarial attacks [13]. For instance, Cheng et al. [6] showed that the multi-modal auto-driving system can be undermined by a single-modal attack that only aims at the camera modality, which is considered less expensive to compromise. Those attacks cause severe security and safety concerns for the deployment of multi-modal models in various real-world applications such as autonomous driving [11, 30, 35, 44, 48]. In our work, we consider a general attack, where an attacker could arbitrarily add (or delete or modify) a certain number of basic elements to each modality. For instance, when each basic element of a modality represents a pixel, an attacker could arbitrarily manipulate (e.g., modify) some pixel values for that modality.

### 2.2. Existing Defenses

Defenses against adversarial attacks can be categorized into *empirical defenses* and *certified defenses*. Empirical defenses [25, 34, 40, 45, 49, 52, 56] cannot provide formal robustness guarantees under arbitrary attacks. Multiple works [4, 5, 46] have shown that they can be bypassed by more advanced attacks. Existing certified defenses [7, 10, 19, 22, 27, 28, 32, 38, 53, 55, 57, 60, 62] against adversarial attacks all focus on unimodal model whose input is only from a single modality. Among those defenses, randomized ablation [21, 28] achieves state-of-the-art certified robustness guarantee when an attacker could arbitrarily modify a certain number of basic elements to the input. Our experimental results show that randomized ablation achieves sub-optimal provable robustness guarantees when extended to multi-modal models. This is because when the attacker introduces perturbations with $l_0$ bounds across all modalities, the space of possible perturbed multi-modal inputs cannot be straightforwardly formulated as a $l_0$ ball.

We note that all the previously discussed certified defenses [7, 10, 19, 27, 28, 53, 57, 60] are model-agnostic and scalable to large models. Another family of certified defenses [14, 23, 51] proposed to derive the certified robustness guarantee of an unimodal model by conducting a layer-by-layer analysis. In general, those methods cannot be applied to general models and are not scalable to large neural networks.

## 3. Problem Formulation

We first introduce the threat model and then formally define certified defense against adversarial attacks to classification and segmentation tasks.

### 3.1. Threat Model

We discuss the threat model from the perspective of the attacker's goals, background knowledge, and capabilities.

**Attacker's goals.** Given a multi-modal input and a multi-modal model, an attacker aims to adversarially perturb the multi-modal input such that the multi-modal model makes incorrect predictions for the perturbed multi-modal input.

**Attacker's background knowledge and capabilities.** As we focus on the certified defense, we assume the attacker has full knowledge about about the multi-modal model, including its architecture and parameters. We consider a strong attack to multi-modal models. In particular, given a multi-modal input, an attacker could simultaneously manipulate all modalities of the input [45, 61]. As a result, a multi-modal makes incorrect predictions for the perturbed multi-modal input. For example, to attack an auto-driving system, the attacker can add adversarial perturbation to both the depth image (captured by a LiDAR depth sensor) and the RGB image (captured by a camera) to lower the prediction quality.

Formally, we denote a multi-modal input as $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_T)$, where $\mathbf{m}_i$ represents a group of elements of the $i$-th modality, the attacker could arbitrarily add (or delete or modify) at most $r_i$ elements to $\mathbf{m}_i$. For instance, when $\mathbf{m}_i$ represents an image, an attacker could arbitrarily change $r_i$ pixel values.

We use $\mathbf{M}' = (\mathbf{m}'_1, \mathbf{m}'_2, \cdots, \mathbf{m}'_T)$ to denote the adversarial input. Without loss of generality, every modality can be rewritten as a list of it's basic elements. For example, an image (e.g., RGB image) can be written as a list of pixels, and an audio can be written as a list of audio frames. Therefore, we can denote $\mathbf{m}_i$ as a composition of basic elements denoted by $[m_i^1, m_i^2, \cdots, m_i^{n_i}]$, where $m_i^j$ represents the $j$-th basic element in the $i$-th modality, and $n_i$ represents the total number of basic elements in the $i$-th modality. We denote the number of basic elements in each modality after the attack as $n'_1, n'_2, \ldots, n'_T$, respectively. For the image modality, we know the number of basic elements (pixels) is fixed. However, for some other modalities like audio, the attacker is able to change the number of basic elements (e.g., audio frames) via addition or deletion.

Hence, we define three kinds of attacks for each modality: modification attack, addition attack, and deletion attack. We use $\mathcal{S}(\mathbf{m}_i, r_i)$ to denote the set of all possible $\mathbf{m}'_i$ when an attacker could add (or delete or modify) at most $r_i$ basic elements in $\mathbf{m}_i$. For simplicity, we use $\mathbf{R} = (r_1, r_2, \ldots, r_T)$ to denote the added (or deleted or modified) basic elements to all modalities. Then we use $\mathcal{S}(\mathbf{M}, \mathbf{R}) = \mathcal{S}(\mathbf{m}_1, r_1) \times \mathcal{S}(\mathbf{m}_2, r_2) \ldots \times \mathcal{S}(\mathbf{m}_T, r_T)$ to denote the set of all possible adversarial inputs $\mathbf{M}' = (\mathbf{m}'_1, \mathbf{m}'_2, \cdots, \mathbf{m}'_T)$.

### 3.2. Certifiably Robust Multi-modal Prediction

For classification tasks, suppose we have a multi-modal classifier $G$. Given a test sample $(\mathbf{M}, y)$, where $y$ is the ground truth label, we say $G$ is *certifiably stable* for $\mathbf{M}$ if the predicted label remains unchanged under attack:

$$G(\mathbf{M}) = G(\mathbf{M}'), \forall \mathbf{M}' \in \mathcal{S}(\mathbf{M}, \mathbf{R}). \tag{1}$$

If this unchanged label is the ground-truth label of $\mathbf{M}$, i.e., $G(\mathbf{M}) = y$, then we say the classifier $G$ is *certifiably robust* for this test sample.

For segmentation tasks, without loss of generality, we assume the multi-modal model outputs the segmentation result for one of the input modalities (denoted by $\mathbf{m}_o$) with $n_o$ basic elements (e.g., pixels). Then the output contains $n_o$ labels. For example, if RGB image is one of the input modalities, the output can be a segmentation of this RGB image, which contains a label for each pixel in the RGB image. Unless otherwise mentioned, we assume that the attacker performs modification attacks on $\mathbf{m}_o$ (please refer to Appendix C for deletion and addition attacks on $\mathbf{m}_o$).

We can think of the multi-modal segmentation model $G$ as composed of multiple classifiers denoted by $G_1, G_2, \ldots, G_{n_o}$. Each classifier $G_j$ predicts a label $G_j(\mathbf{M})$ for $m_o^j$ (the $j$-th basic element of $\mathbf{m}_o$). The ground truth $y$ also includes $n_o$ labels, denoted by $y_1, y_2, \ldots, y_{n_o}$. We use $G_j(\mathbf{M}')$ to denote the predicted label for $m_o^j$ after the attack. We say $G_j$ is *certifiably stable* for a basic element (e.g., a pixel) $m_o^j$ if:

$$G_j(\mathbf{M}) = G_j(\mathbf{M}'), \forall \mathbf{M}' \in \mathcal{S}(\mathbf{M}, \mathbf{R}), \tag{2}$$

which means the predicted label for the the $j$-th basic element of $\mathbf{m}_o$ remains unchanged under attack. If $G_j(\mathbf{M}) = y_j$, then we term $G_j$ as *certifiably robust* for $m_o^j$.

By deriving a lower bound on the number of basic elements whose predictions are certifiably robust, we can guarantee the segmentation quality for a test sample, measured via metrics such as Certified Pixel Accuracy, Certified F-score, or Certified IoU.

## 4. Our Design

### 4.1. Independent Sub-sampling

In this section, we will first outline a universal sub-sampling method [21, 28, 32], and then demonstrate its application across various multi-modal tasks.

**Sub-sampling Strategy.** We repeatedly randomly sub-sample $k_i$ basic elements (e.g., pixels) from the $i$-th modality $\mathbf{m}_i = [m_i^1, m_i^2, \cdots, m_i^{n_i}]$ without replacement. For simplicity, we use $\mathcal{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)$ to denote the randomly sampled multi-modal input. Thus, we have $|\mathbf{z}_i| = k_i$

for all $i = 1, 2, \ldots, T$. This sampling strategy exhibits versatility by being applicable across various modalities and tasks. It can be applied for classification tasks, e.g., emotion recognition. And it can also be employed for segmentation tasks, e.g., road segmentation. Figure 9 in Appendix provides a visualization of this sub-sampling method.

Next, we first apply this sampling strategy to build an ensemble classifier for classification tasks.

### 4.2. Certify Multi-modal Classification

**Ensemble Classifier.** Given a testing input $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_T)$, we use $\mathcal{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)$ to denote the randomly sub-sampled multi-modal input. We denote the multi-modal model by $g$. For simplicity, we use $g(\mathcal{Z})$ and $y$ to denote the predicted label and the true label. As $\mathcal{Z}$ is randomly sub-sampled, $g(\mathcal{Z})$ is also random. Given an arbitrary label $l \in \{1, 2, \cdots, C\}$ ($C$ is the total number of classes), we use $p_l$ to denote the probability that the predicted label $g(\mathcal{Z})$ is $l$. Formally, we have $p_l = \Pr(l = g(\mathcal{Z}))$. We call $p_l$ *label probability*. In practice, it is computationally expensive to calculate the exact label probabilities. Following [10, 19, 28], we use Monte Carlo sampling to estimate a lower bound or upper bound of $p_l$, denoted as $\underline{p_l}$ and $\overline{p_l}$ respectively. This is achieved by randomly sample $N$ ablated inputs from the distribution $\mathcal{Z}$, represented as $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_N$, and then count the label frequency $N_l = \sum_{i=1}^{N} \mathbb{I}(g(\mathbf{Z}_i) = l)$ for each label $l$. Our ensemble classifier $G$ then predicts the label with the largest frequency $N_l$. For simplicity, we denote this label by $A$ and use $\underline{p_A}$ to represent $A$'s label probability lower bound. We define the runner-up label $B$ as the label with the second highest label frequency, i.e., $B = \text{argmax}_{l \neq A} N_l$. We present our certification result below:

**Theorem 1** (Certification for classification)**.** *Suppose we have a multi-modal test input $\mathbf{M}$ and a base multi-modal classifier $g$. Our ensemble classifier $G$ is as defined as above. We denote $A = G(\mathbf{M})$ and use $\underline{p_A}$ to denote the label probability lower bound for the label $A$. We use $B$ to denote the runner-up class and use $\overline{p_B}$ to denote the label probability upper bound for the label $B$. We define $\delta_l = \underline{p_A} - \frac{\lfloor \underline{p_A} \prod_{i=1}^{T} \binom{n_i}{k_i} \rfloor}{\prod_{i=1}^{T} \binom{n_i}{k_i}}$ and $\delta_u = \frac{\lceil \overline{p_B} \prod_{i=1}^{T} \binom{n_i}{k_i} \rceil}{\prod_{i=1}^{T} \binom{n_i}{k_i}} - \overline{p_B}$. Given a perturbation size $\mathbf{R} = (r_1, r_2, \ldots, r_T)$, we have the following:*

$$G(\mathbf{M}) = G(\mathbf{M}'), \forall \mathbf{M}' \in \mathcal{S}(\mathbf{M}, \mathbf{R}) \tag{3}$$

*if:*

$$\frac{\prod_{i=1}^{T} \binom{n_i}{k_i}}{\prod_{i=1}^{T} \binom{n'_i}{k_i}} (\underline{p_A} - \delta_l - 1 + \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i}{k_i}}) \tag{4}$$

$$\geq \frac{\prod_{i=1}^{T} \binom{n_i}{k_i}}{\prod_{i=1}^{T} \binom{n'_i}{k_i}} (\overline{p_B} + \delta_u) + 1 - \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n'_i}{k_i}} \tag{5}$$

*where $e_i = n_i - r_i$ and $n'_i = n_i$ for modification attack; $e_i = n_i$ and $n'_i = n_i + r_i$ for addition attack; $e_i = n_i - r_i$ and $n'_i = n_i - r_i$ for deletion attack, where $i = 1, 2, \ldots, T$ is the modality index.*

*Proof.* Please refer to Appendix A. □

**Computing $\underline{p_B}$ and $\overline{p_A}$.** Following [10, 19, 20], we apply Monte Carlo sampling to approximate $\underline{p_B}$ and $\overline{p_A}$. We first randomly sub-sample $N$ multi-modal inputs from the test input $\mathbf{M}$, and we denote these ablated inputs as $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_N$. We denote the number of sub-sampled inputs that predicts for the label $l$ as $N_l$, i.e., $N_l = \sum_{i=1}^{N} \mathbb{I}(g(\mathbf{Z}_i) = l)$. Then, the frequency $N_l$ of any label $l$ follows a binomial distribution. Therefore, we can apply Clopper-Pearson [9] based method to estimate $\underline{p_B}$ and $\overline{p_A}$ with predefined confidence level $1 - \alpha$:

$$\underline{p_A} = Beta(\frac{\alpha}{C}; N_A, N - N_A + 1) \tag{6}$$

$$\overline{p_B} = Beta(1 - \frac{\alpha}{C}; N_B, N - N_B + 1), \tag{7}$$

where $A$ represents the predicted label, i.e., $A = \text{argmax}_l N_l$, and $B$ represents the runner-up label, i.e., $\text{argmax}_{l \neq A} N_l$. $Beta(\beta; \lambda, \theta)$ calculates the $\beta$-th quantile of the Beta distribution given shape parameters $\lambda$ and $\theta$. We divide $\alpha$ by the number of classes because we estimate bounds for $C$ classes simultaneously [20]. By Bonferroni correction, if we use $1 - \alpha/C$ as the confidence level to estimate each bound, then the overall confidence level for the $C$ classes is at least $1 - \alpha$.

### 4.3. Certify Multi-modal Segmentation

In this section, we extend our certification method for classification tasks to certify multi-model segmentation tasks. Segmentation tasks are essentially a variant of classification since each basic element (e.g., a pixel) in one of the input modalities (e.g., an image) is assigned a label. We denote this input modality as $\mathbf{m}_o$, and denote the $j$-th basic element of $\mathbf{m}_o$ as $m_o^j$. Suppose $\mathbf{m}_o$ has $n_o$ basic elements. If we naively apply union bound, certifying the test input with overall confidence level $1 - \alpha$ requires certifying each basic element with confidence level $1 - \frac{\alpha}{n_o}$, which becomes hard when $n_o$ grows large. To maximize the number of certified basic elements, Fischer et al. [12] utilized the Holm–Bonferroni method [17], originally designed for *Multiple Hypothesis Testing*. Specifically, this method tends to certify basic elements with confident predictions, while abstaining ambiguous basic elements. Furthermore, this method guarantees that the probability of mistakenly reporting at least one non-certifiably-stable basic element as certified is limited at $\alpha$. In this work, we adapt the approach from Fischer et al. [12] to multi-modal scenarios.

**Ensemble Classifiers for Segmentation.** Given a testing input $\mathbf{M}$, we use $\mathcal{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)$ to denote the

randomly sub-sampled input. The base multi-modal segmentation model can be seen as a composition of multi-modal classifiers $g_1, g_2, \ldots, g_{n_o}$, where $g_j$ predicts a label for the basic element $m_o^j$. We use $g_j(\mathcal{Z})$ to denote the predicted label for $m_o^j$. We randomly sample $N$ ablated inputs from the distribution $\mathcal{Z}$, and represent them as $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_N$. For each basic element $m_o^j$ and each label $l$, we count the label frequency $N_l^j = \sum_{i=1}^{N} \mathbb{I}(g_j(\mathbf{Z}_i) = l)$. The ensemble classifier for $m_o^j$ (denoted by $G_j$) then predicts the the label $l$ with the highest label frequency $N_l^j$, i.e., $G_j(\mathbf{M}) = \text{argmax}_l N_l^j$. We say $G_j$ is *certifiably stable* for $m_o^j$ if the predicted label of $G_j$ for $m_o^j$ remains unchanged under attack, i.e., $G_j(\mathbf{M}) = G_j(\mathbf{M}'), \forall \mathbf{M}' \in \mathcal{S}(\mathbf{M}, \mathbf{R})$. Next, we discuss how to certify as many basic elements as possible given that the possibility of mistakenly certifying a non-certifiably-stable basic element is at most $\alpha$.

**Calculate a Confidence Level for Each Basic Element.** For each basic element $m_o^j$, we denote the number of ablated inputs that predicts the label $l$ for this component as $N_l^j$. We denote the total number of ablated inputs as $N$. We define:

$$\underline{p_A}(\alpha_j) = Beta(\frac{\alpha_j}{C}; N_A^j, N - N_A^j + 1) \tag{8}$$

$$\overline{p_B}(\alpha_j) = Beta(1 - \frac{\alpha_j}{C}; N_B^j, N - N_B^j + 1), \tag{9}$$

where $A$ represents the predicted label for this basic element, i.e., $\text{argmax}_l N_l^j$, and $B$ represents the runner-up label for this component, i.e., $\text{argmax}_{l \neq A} N_l^j$. Then we define:

$$\alpha_j^* = \min_{\alpha_j} \tag{10}$$

$$s.t., \frac{\prod_{i=1}^{T} \binom{n_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}} (\underline{p_A}(\alpha_j) - \delta_l - 1 + \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i}{k_i}}) \tag{11}$$

$$\geq \frac{\prod_{i=1}^{T} \binom{n_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}} (\overline{p_B}(\alpha_j) + \delta_u) + 1 - \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}}, \tag{12}$$

where $n_i$, $n_i'$, $e_i$, $k_i$, $\delta_l$ and $\delta_u$ are defined as in Theorem 1. Then with probability at least $1 - \alpha_j^*$, the basic element $m_o^j$ is certifiably stable (the output label of this basic element cannot be changed by the attacker) according to Theorem 1. In practice, we calculate $\alpha_j^*$ by binary search. If such an $\alpha_j^*$ does not exist, the binary search algorithm returns 1 instead.

**Apply Holm-Bonferroni method.** Using the computed values of $\alpha_j^*$, we employ the Holm-Bonferroni method [17] to determine the basic elements eligible for certification. This method maximizes the number of certified basic elements, while at the same time ensures that the possibility of mistakenly certifying a non-certifiably-stable basic element remains within the limit of $\alpha$. Specifically, we have two steps:

- Step 1: We order $\alpha_j^*$-values ($j = 1, 2, \cdots, n_o$) in ascending order so that we have $\alpha_{(1)}^* \leq \alpha_{(2)}^* \leq \cdots \leq \alpha_{(n_o)}^*$.
- Step 2: Calculate $L = \min\{j : \alpha_{(j)}^* > \frac{\alpha}{n_o + 1 - j}\}$.

We report all basic elements $m_o^j$ for which $\alpha_j^* < \alpha_{(L)}^*$ as certifiably stable (the output labels of these basic elements cannot be changed by the attacker), and the predictions for other basic elements are abstained. In Section 5, we derive certified metrics, e.g., Certified Pixel Accuracy, from these certifiably stable basic elements.

## 5. Evaluation

In this section, we demonstrate the effectiveness of our method on multi-modal emotion recognition task and multi-modal road segmentation task.

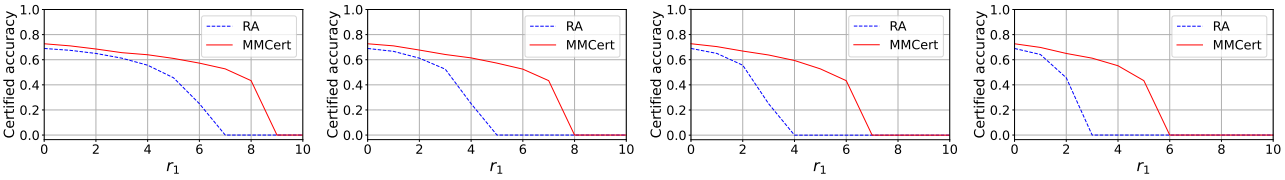
### 5.1. Experimental Setup

**Datasets.** We use the following benchmark datasets in our evaluation: RAVDESS [33] for the multi-modal emotion recognition task and KITTI Road [1] for the multi-modal road segmentation task. Details of the datasets can be found in Appendix B.

**Models.** For the multi-modal emotion recognition task, we follow the pipeline proposed by [8]. Specifically, we utilize EfficientFace [63] (a recently proposed facial expression recognition architecture) to extract features from image frames, and use 1D convolutional layers to extract features from audio frames. Then we use intermediate attention-based fusion [8] to combine features extracted from these two modalities. This fusion method ensures that features that are consistent between both modalities have the most significant impact on the final prediction.

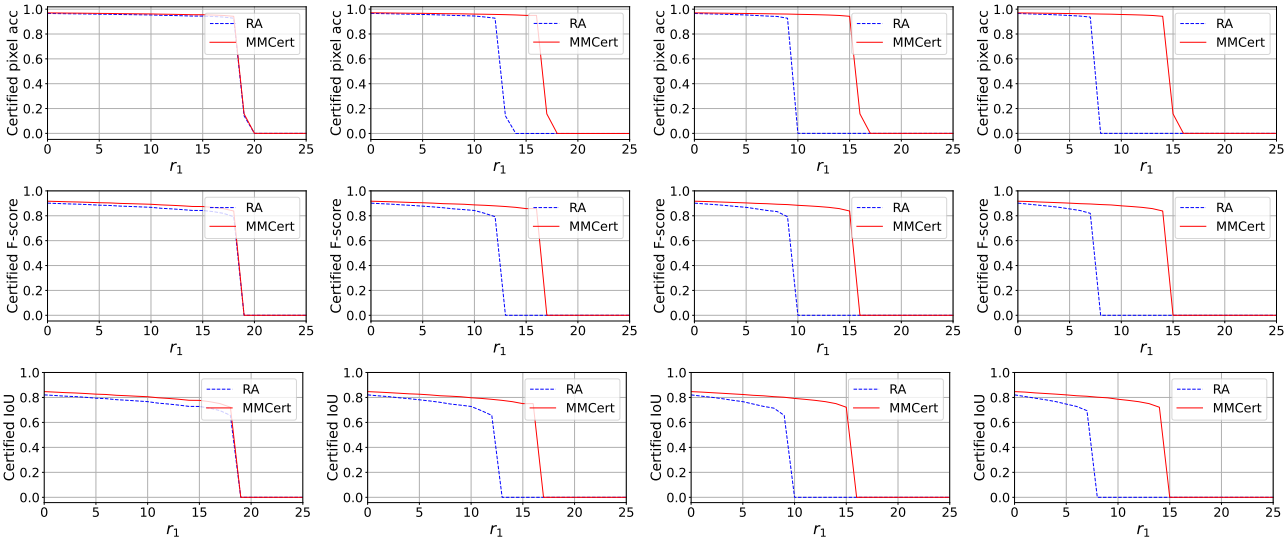
As for the multi-modal road segmentation task, we apply SNE-RoadSeg [11], which is capable of merging features from both RGB images and depth images for road segmentation. Specifically, this method first computes surface normal information from depth images, and then employs a data-fusion CNN architecture to fuse features from both RGB images and the inferred surface normal information for accurate prediction.

We note that if we directly use the original training recipes for these models, we get low prediction accuracy for randomly sub-sampled testing inputs, and the certified robustness of MMCert and randomized ablation [29] would be low as both rely on predicting ablated inputs. In response, we perform data augmentation by randomly ablate training inputs, such that the distribution of training data can match that of testing data. We note that this is standard practice for randomized smoothing-based certification methods [10]. For MMCert, we independently sub-sample between 0% and 5% of basic elements from each modality, ablating the rest. For randomized ablation [29], we randomly sample between 0% and 5% of basic elements collectively from the two modalities to keep and ablate the remaining elements. Specifically, we initially merge the basic




(a) $r_2 = r_1$ (b) $r_2 = 2r_1$ (c) $r_2 = 3r_1$ (d) $r_2 = 4r_1$

Figure 1. Compare our MMCert with randomized ablation on RAVDESS Dataset.




(a) $r_2 = r_1$ (b) $r_2 = 2r_1$ (c) $r_2 = 3r_1$ (d) $r_2 = 4r_1$

Figure 2. Compare our MMCert with randomized ablation on KITTI Road Dataset. Certified Pixel Accuracy (first row), Certified F-score (second row) and Certified IoU (third row) are considered.

elements of both modalities into one list. After sampling and ablating, we then split the modified list back into two separate modalities.

**Compared Method.** We compare our method with randomized ablation [28], which is the state-of-the-art certification method for $l_0$ attacks on a single image.

We adapt randomized ablation to multi-modal models by combining the sets of basic elements from each modality. Given the original input $(\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_T)$, where $\mathbf{m}_i = [m_i^1, m_i^2, \ldots, m_i^{n_i}]$ , we combine all modalities to get $\mathbf{m} = [m_1^1, m_1^2, \ldots, m_1^{n_1}, \ldots, m_T^1, m_T^2, \ldots, m_T^{n_T}]$. We denote the size of $\mathbf{m}$ as $n = \sum_{i=1}^{T} n_i$. Then we randomly sample $k$ elements from $\mathbf{m}$ without replacement to get a subset $\mathbf{z} \subseteq \mathbf{m}$. Finally, we divide $\mathbf{z}$ back to $T$ modalities, i.e., $\mathbf{z}_i = \mathbf{z} \cap \mathbf{m}_i$, and $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)$ is the randomly ablated input. We make the final prediction by taking the majority vote of all ablated multi-modal inputs.

For multi-modal classification tasks, we use the same certification process for randomized ablation as in the original paper [28]. For multi-modal segmentation tasks, we follow the same certification process as described in Section 4.3 as the original work [28] only considered classification tasks. The only difference is that we define $\alpha_j^*$ as:

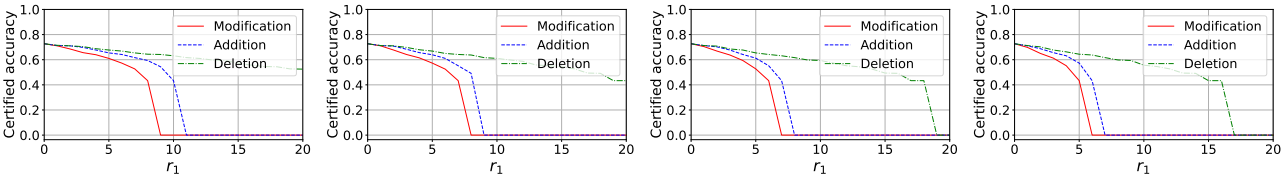
$$\alpha_j^* = \min_{\alpha_j} \tag{13}$$

$$s.t., \quad \underline{p_A}(\alpha_j) - 1 + \frac{\binom{n-\sum_{i=1}^{T} r_i}{k}}{\binom{n}{k}} \geq \overline{p_B}(\alpha_j) + 1 - \frac{\binom{n-\sum_{i=1}^{T} r_i}{k}}{\binom{n}{k}}. \tag{14}$$

It is worth noting that in this context, $r_i$ represent the maximum number of modified basic elements in the $i$-th modality. The original work did not take into account addition and deletion attacks, as [28] focuses on image domain.

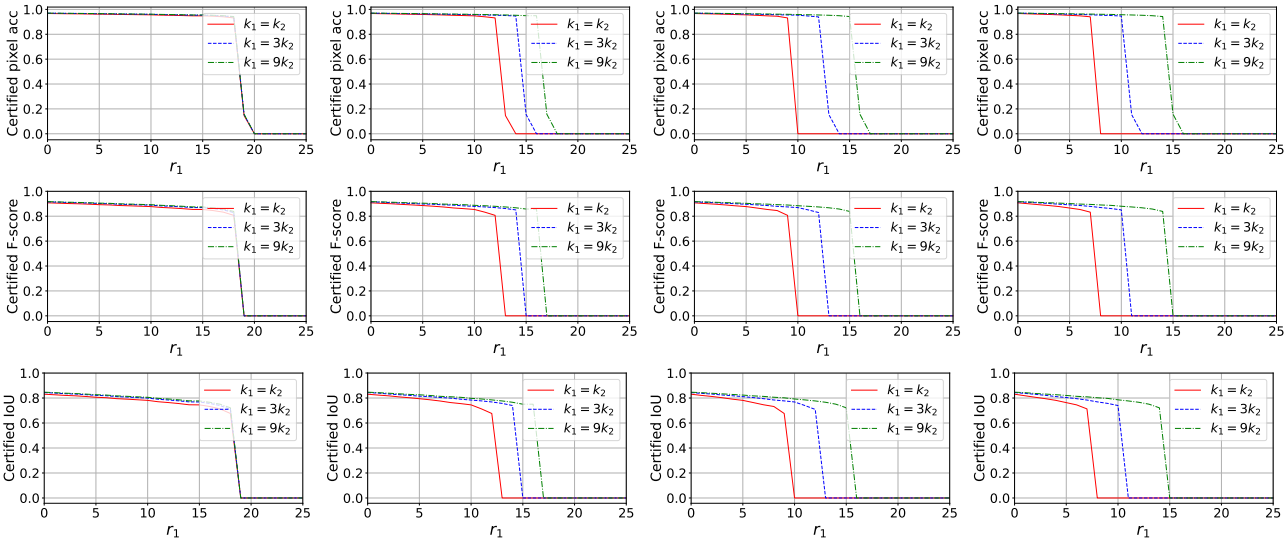
**Parameter Settings.** By default, we focus on modification attacks, where $r_i$ denote the maximum basic elements that can be modified by the attacker in $i$th modality.

For the multi-modal emotion recognition task, the visual modality contains 108 image frames, while the audio modality contains 79,380 audio frames. Without loss of generality, we denote the maximum number of modified image frames as $r_1$ and the maximum number of modified audio frames as $r_2$. The default setting is that the attacker can modify equal or more audio frames than image frames. That is, we let $r_2 = \hat{c} \cdot r_1$, for $\hat{c} = 1, 2, 3, 4$. We set $k_1 = 5$




(a) $r_2 = r_1$ (b) $r_2 = 2r_1$ (c) $r_2 = 3r_1$ (d) $r_2 = 4r_1$

Figure 3. Compare different attack types on RAVDESS Dataset.




(a) $r_2 = r_1$ (b) $r_2 = 2r_1$ (c) $r_2 = 3r_1$ (d) $r_2 = 4r_1$

Figure 4. Impact of the ratio between $k_1$ and $k_2$. Certified Pixel Accuracy (first row), Certified F-score (second row) and Certified IoU (third row) are considered.

and $k_2 = 1,000$. For randomized ablation, $k$ is set to 3,000 such that, when there is no attack ($r_1 = r_2 = 0$), the accuracy of randomized ablation is similar to our MMCert. In Appendix D, we show the case where an attacker can modify more image frames than audio frames ($r_1 > r_2$).

For the multi-modal road segmentation task, the first modality is a RGB image that consists of $375 \times 1,242$ pixels, where each pixel has three channels (representing the three primary colors), while the second modality is a depth image that has the same number of pixels, but each pixel has a single channel for depth. The default setting is that the attacker can modify equal or more pixels from the depth image than pixels from the RGB image. Specifically, we test for $r_2 = \hat{c} \cdot r_1$ where $\hat{c} = 1, 2, 3, 4$. For our MMCert, we set $k_1 = 9,000$ and $k_2 = 1,000$. Regarding randomized ablation, we set the total number of retained pixels $k$ to 10,000 such when there is no attack ($r_1$ = $r_2$ = 0), the accuracy of randomized ablation is similar to our MMCert. In Appendix D, we show the case where the attacker can change more pixels from the RGB image than pixels from the depth image ($r_1 > r_2$).

For Monte Carlo sampling, we set $N = 100$ and $\alpha = 0.001$ for all experiments.

**Evaluation Metrics.** We use *Certified Accuracy* as the evaluation metric for the multi-modal emotion recognition task, and use *Certified Pixel Accuracy*, *Certified F-score*, and *Certified IoU* as the evaluation metrics for the multi-modal road segmentation task.

- **Certified Accuracy.** Certified Accuracy is defined as the fraction of testing inputs whose predicted labels are not only correct but also verified to be unchanged by an attacker, i.e., certifiably stable. A testing sample for a multi-modal model can be represented as $(\mathbf{M}, y) \in D_{test}$, where $D_{test}$ is the testing dataset. $\mathbf{M}$ is the multi-modal test input and $y$ is the ground truth label. We use $G$ to denote the multi-modal classifier. Then we can define Certified Accuracy as:

$$\frac{\sum_{(\mathbf{M},y)\in\mathcal{D}_{test}} \mathbb{I}(IsStable(\mathbf{M}) \wedge G(\mathbf{M}) = y)}{|\mathcal{D}_{test}|}. \quad (15)$$

  $IsStable(\mathbf{M})$ is true if and only if for all $\mathbf{M}' \in \mathcal{S}(\mathbf{M}, \mathbf{R})$, we have $G(\mathbf{M}) = G(\mathbf{M}')$. $\mathbb{I}$ is the indicator function, and $|\mathcal{D}_{test}|$ is the total number of testing inputs in $\mathcal{D}_{test}$.

- **Certified Pixel Accuracy (or F-score or IoU).** Here

we consider these certified metrics for the purpose of freespace detection [11]. For general purposed segmentation tasks, mean values over different classes should be considered. The Certified Pixel Accuracy (or F-score or IoU) is defined as the average Pixel Accuracy (or F-score or IoU) lower bound of testing inputs under a given adversarial perturbation space $\mathbf{R}$. We use $j \in [n_o]$ to denote the index of a basic element of the segmented input modality $\mathbf{m}_o$. We define:

$$TP = |\{j : (G_j(\mathbf{M}) = y_j = 1) \wedge IsStable(\mathbf{M}, j)\}|,$$

$$TN = |\{j : (G_j(\mathbf{M}) = y_j = 0) \wedge IsStable(\mathbf{M}, j)\}|,$$

$$FP = |\{j : G_j(\mathbf{M}) = 1\}| - TP, \text{and}$$

$$FN = |\{j : G_j(\mathbf{M}) = 0\}| - TN,$$

where label 1 represents freespace and label 0 represents non-freespace. $IsStable(\mathbf{M}, j)$ is true if and only if the predicted label of the $j$th basic element of $\mathbf{m}_o$ cannot be changed by the attacker, i.e., $G_j(\mathbf{M}) = G_j(\mathbf{M}'), \forall \mathbf{M}' \in \mathcal{S}(\mathbf{M}, \mathbf{R})$. Then for an individual test sample, we have Certified Pixel Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$, Certified F-score $= \frac{2TP^2}{2TP^2+TP(FP+FN)}$, and Certified IoU $= \frac{TP}{TP+FP+FN}$. To obtain the final metrics, we compute the average of these values across all test samples.

### 5.2. Experimental Results

In this section, we first compare our method with an existing state-of-the-art method, followed by an analysis of the impact of hyper-parameters on MMCert. Then, we show the performance of our method on attack types other than modification attack, i.e., addition and deletion attacks.

**Our MMCert Outperforms Existing State-of-the-Art Method.** Figure 1 and Figure 2 show the comparison result between our MMCert and randomized ablation [28], which is the state-of-the-art certified defense against $l_0$ attacks. We can see that our MMCert consistently outperforms randomized ablation on both tasks, for all combinations of $r_1$ and $r_2$ . For example, Figure 1 shows that on the RAVDESS dataset, when $r_1 = r_2 = 8$ (the attacker can modify 8 frames in both visual and audio modalities), our MMCert can guarantee correct predictions for more than $40\%$ of the test samples, while randomized ablation can guarantee $0\%$ of the test samples.

Our method is more effective than randomized ablation because of two reasons. First, our method provides an adaptive selection of $k_1$ and $k_2$ to control the fraction of sub-sampled basic elements, i.e., $\frac{k_1}{n_1}$ and $\frac{k_2}{n_2}$, of the two modalities. In contrast, for randomized ablation, the sub-sampled fractions for both modalities are the identical on average, i.e., $\frac{k}{n}$. This means that randomized ablation is essentially a special case of our MMCert. Secondly, our MMCert is more stable than randomized ablation during both training and testing phases. In our method, the count of sub-sampled basic elements remains constant at $k_1$ and $k_2$ for each modality. Meanwhile, in randomized ablation, this count, adding up to $k$, fluctuates. As a result, our method's sub-sampled input space is smaller than that of randomized ablation, enhancing stability.

**Impact of $k_1$ and $k_2$.** Here we study the impact of $k_1$ and $k_2$ on the performance of our MMCert. To simplify the analysis, we perform the experiment on KITTI Road Dataset such that we have $n_1 = n_2$. This setup allows a direct comparison of the attacker's capability across two modalities using $r_1$ and $r_2$. We keep the sum of $k_1$ and $k_2$ constant at 10,000 but vary their ratio. Three specific ratios were tested: $k_1 = k_2$, $k_1 = 3k_2$, and $k_1 = 9k_2$. The results are presented in Figure 4. We observe that with $r_2 = r_1$ (indicating similar attack capabilities on both modalities), different ratios of $k_1$ and $k_2$ have similar performance outcomes. However, for $r_2 > r_1$ (where the attacker has more attack capability on the second modality), strategies with a larger $k_1/k_2$ ratio demonstrated better robustness. For example, if $r_2 > r_1$, the $k_2 = 9k_1$ sub-sampling strategy consistently outperforms $k_2 = 3k_1$, with this advantage magnifying as $r_2/r_1$ increased.. Therefore, in practice, it is advantageous to sub-sample fewer basic elements from the modality with higher attack capability and sub-sample more basic elements from the modality with lower attack capability, provided this doesn't compromise the utility (accuracy when there is no attack).

**Different Attack Types.** We previously focused on modification attacks, where the attacker modifies at most $r_1$ and $r_2$ basic elements for the two respective modalities. Our method also allows the attacker to add or delete basic elements from each modality. Here we do experiments in scenarios where the attacker can add (or delete) at most $r_1$ and $r_2$ basic elements respectively for the two modalities, and compare with the modification attack scenario. The results are shown in Figure 3. We can see that modification attack is the strongest attack type. For example, when $r_1$ and $r_2$ are both 10, modification attack brings the certified accuracy down to 0. In contrast, the addition attack maintains a certified accuracy greater than 0.4, and the deletion attack maintains a certified accuracy greater than 0.6.

## 6. Conclusion

In this work, we propose MMCert, the first certified defense against adversarial attacks for multi-modal models. Our experimental results show that MMCert significantly improves the certified robustness guarantees by leveraging a modality-independent sub-sampling strategy.

**Acknowledgements.** We thank the anonymous reviewers for their valuable feedback on our paper, which significantly improved the quality of our work.

## References


[1] KITTI Road Dataset. https://www.cvlibs.net/datasets/kitti/eval_road.php. Accessed: 2023-10-01. 5, 14

[2] Multi-modal Emotion Recognition Implementation. https://github.com/katerynaCh/multimodal-emotion-recognition. Accessed: 2023-10-01. 13

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2

[4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 1, 2

[5] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017. 1, 2

[6] Zhiyuan Cheng, Hongjun Choi, James Liang, Shiwei Feng, Guanhong Tao, Dongfang Liu, Michael Zuzak, and Xiangyu Zhang. Fusion is not enough: Single-modal attacks to compromise fusion models in autonomous driving. *arXiv preprint arXiv:2304.14614*, 2023. 1, 2

[7] Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified defenses for adversarial patches. *arXiv preprint arXiv:2003.06693*, 2020. 1, 2

[8] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Self-attention fusion for audiovisual emotion recognition with incomplete data. In *ICPR*. IEEE, 2022. 1, 2, 5, 13

[9] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 1934. 4

[10] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019. 1, 2, 4, 5, 12

[11] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *ECCV*, 2020. 1, 2, 5, 8

[12] Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing. In *ICML*, 2021. 4

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[14] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018. 1, 2

[15] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2019. 1

[16] Ehtesham Hassan, Yasser Khalil, and Imtiaz Ahmad. Learning feature fusion in deep learning-based object detector. *Journal of Engineering*, 2020. 2

[17] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 1979. 4, 5

[18] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, 2020. 2

[19] Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *ICLR*, 2020. 1, 2, 4

[20] Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic certified robustness of bagging against data poisoning attacks. In *AAAI*, 2021. 2, 4, 12

[21] Jinyuan Jia, Binghui Wang, Xiaoyu Cao, Hongbin Liu, and Neil Zhenqiang Gong. Almost tight l0-norm certified robustness of top-k predictions against adversarial perturbations. In *ICLR*, 2021. 2, 3

[22] Jinyuan Jia, Wenjie Qu, and Neil Gong. Multiguard: Provably robust multi-label classification against adversarial examples. *Advances in Neural Information Processing Systems*, 35:10150–10163, 2022. 2

[23] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*, 2017. 2

[24] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019. 1, 2

[25] Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. *NeurIPS*, 2019. 1, 2

[26] Ayush Kumar and Jithendra Vepa. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP*. IEEE, 2020. 2

[27] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019. 1, 2

[28] Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. *CoRR*, abs/1911.09272, 2019. 1, 2, 3, 4, 6, 8

[29] Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *AAAI*, number 04, 2020. 5

[30] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *CVPR*, 2022. 1, 2

[31] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *CVPR*, 2022. 2

[32] Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Pointguard: Provably robust 3d point cloud classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6186–6195, 2021. 2, 3

[33] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 2018. 5, 13

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv*, 2017. 1, 2

[35] Oskar Natan and Jun Miura. End-to-end autonomous driving with semantic depth cloud mapping and multi-agent. *IEEE Transactions on Intelligent Vehicles*, 2022. 1, 2

[36] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London.*, 1933. 2, 12

[37] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015. 1

[38] Hengzhi Pei, Jinyuan Jia, Wenbo Guo, Bo Li, and Dawn Song. Textguard: Provable defense against backdoor attacks on text classification. In *NDSS*, 2024. 2

[39] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multimodal fusion transformer for end-to-end autonomous driving. In *CVPR*, 2021. 1, 2

[40] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *NeurIPS*, 2019. 1, 2

[41] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *CVPR*, 2019. 1, 2

[42] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016. 1

[43] Junjie Shen, Jun Yeon Won, Zeyuan Chen, and Qi Alfred Chen. Drift with devil: Security of {Multi-Sensor} fusion based localization in {High-Level} autonomous driving under {GPS} spoofing. In *USENIX Security*, 2020. 1, 2

[44] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018. 1, 2

[45] Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *CVPR*, 2021. 1, 2, 3

[46] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *ICML*, 2018. 1, 2

[47] Jörg Wagner, Volker Fischer, Michael Herman, Sven Behnke, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, 2016. 2

[48] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, 2021. 1, 2

[49] Wenjie Wang, Pengfei Tang, Jian Lou, and Li Xiong. Certified robustness to word substitution attack with differential privacy. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. 1, 2

[50] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In *ICCV*, 2023. 1

[51] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*, 2018. 1, 2

[52] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020. 1, 2

[53] Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. {PatchCleanser}: Certifiably robust defense against adversarial patches for any image classifier. In *USENIX Security*, 2022. 1, 2

[54] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *CVPR*, 2018. 1, 2

[55] Han Yang, Binghui Wang, Jinyuan Jia, et al. Graphguard: Provably robust graph classification against adversarial attacks. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[56] Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *CVPR*, 2021. 1, 2

[57] Mao Ye, Chengyue Gong, and Qiang Liu. Safer: A structure-free approach for certified robustness to adversarial word substitutions. *arXiv*, 2020. 1, 2

[58] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 2

[59] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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)*, 2018. 2

[60] Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized [mask]. *Computational Linguistics*, 2023. 1, 2

[61] 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*, 2022. 1, 2, 3

[62] Jinghuai Zhang, Jinyuan Jia, Hongbin Liu, and Neil Zhenqiang Gong. Pointcert: Point cloud classification with deterministic certified robustness guarantees. In *CVPR*, 2023. 2

[63] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *AAAI*, 2021. 5

## A. Proof of Theorem 1

Our proof is extended from previous studies [20]. We first specify notations and then show our proof. Given original multi-modal input pair $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_T)$ and attacked input pair $(\mathbf{m}'_1, \mathbf{m}'_2, \ldots, \mathbf{m}'_T)$, we respectively use $\mathcal{X}$ and $\mathcal{Y}$ to denote the ablated multi-modal input sampled from them without replacement. We use $e_i$ to denote the number of basic elements (e.g., pixels) that are in both $\mathbf{m}_i$ and $\mathbf{m}'_i$, i.e., $e_i = |\mathbf{m}_i \cap \mathbf{m}'_i|$. Moreover, we use $\Upsilon$ to denote the joint space between $\mathcal{X}$ and $\mathcal{Y}$. We use $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_T)$ to denote a variable in the space $\Upsilon$.

We divide the space $\Upsilon$ into the following subspace:

$$\tilde{B} = \{\mathcal{E} | \mathcal{E}_1 \subseteq (\mathbf{m}_1 \cap \mathbf{m}'_1), \mathcal{E}_2 \subseteq (\mathbf{m}_2 \cap \mathbf{m}'_2), \ldots, \mathcal{E}_T \subseteq (\mathbf{m}_T \cap \mathbf{m}'_T)\}, \tag{16}$$

$$\tilde{A} = \{\mathcal{E} | \mathcal{E}_1 \subseteq \mathbf{m}_1, \mathcal{E}_2 \subseteq \mathbf{m}_2, \ldots, \mathcal{E}_T \subseteq \mathbf{m}_T\} - \tilde{B}, \tag{17}$$

$$\tilde{C} = \{\mathcal{E} | \mathcal{E}_1 \subseteq \mathbf{m}'_1, \mathcal{E}_2 \subseteq \mathbf{m}'_2, \ldots, \mathcal{E}_T \subseteq \mathbf{m}'_T\} - \tilde{B}. \tag{18}$$

We present Neyman Pearson Lemma [10, 20, 36] for later use.

**Lemma 1** (Neyman Pearson)**.** *Let $\mathcal{X}$, $\mathcal{Y}$ be two random variables whose probability densities are respectively $Pr(\mathcal{X} = \mathcal{E})$ and $Pr(\mathcal{Y} = \mathcal{E})$, where $\mathcal{E} \in \Upsilon$. Let $Z$ be a random or deterministic functions. where $Z(1|\mathcal{E})$ denotes the probability that $Z(\mathcal{E}) = 1$. Then, we have the following:*

*(1) If $W_1 = \{\mathcal{E} \in \Upsilon : Pr(\mathcal{Y} = \mathcal{E})/Pr(\mathcal{X} = \mathcal{E}) < \mu\}$ and $W_2 = \{\mathcal{E} \in \Upsilon : Pr(\mathcal{Y} = \mathcal{E})/Pr(\mathcal{X} = \mathcal{E}) = \mu\}$ for some $\mu > 0$. Let $S = W_1 \cup W_3$, where $W_3 \subseteq W_2$. If $Pr(Z(\mathcal{X}) = 1) \geq Pr(\mathcal{X} \in S)$, then $Pr(Z(\mathcal{Y}) = 1) \geq Pr(\mathcal{Y} \in S)$.*

*(2) If $W_1 = \{\mathcal{E} \in \Upsilon : Pr(\mathcal{Y} = \mathcal{E})/Pr(\mathcal{X} = \mathcal{E}) > \mu\}$ and $W_2 = \{\mathcal{E} \in \Upsilon : Pr(\mathcal{Y} = \mathcal{E})/Pr(\mathcal{X} = \mathcal{E}) = \mu\}$ for some $\mu > 0$. Let $S = W_1 \cup W_3$, where $W_3 \subseteq W_2$. If $Pr(Z(\mathcal{X}) = 1) \leq Pr(\mathcal{X} \in S)$, then $Pr(Z(\mathcal{Y}) = 1) \leq Pr(\mathcal{Y} \in S)$.*

*Proof.* Let's start by proving part (1). For convenience, we denote the complement of $S$ as $S^c$. With this notation, we have the following:

$$\Pr(Z(\mathcal{Y}) = 1) - \Pr(\mathcal{Y} \in S) \tag{19}$$

$$= \int_{\Upsilon} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} - \int_{S} \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} \tag{20}$$

$$= \int_{S^c} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} + \int_{S} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} - \int_{S} \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} \tag{21}$$

$$= \int_{S^c} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} - \int_{S} (1 - Z(1|\mathcal{E})) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} \tag{22}$$

$$\geq \mu \cdot [\int_{S^c} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E} - \int_{S} (1 - Z(1|\mathcal{E})) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E}] \tag{23}$$

$$= \mu \cdot [\int_{S^c} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E} + \int_{S} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E} - \int_{S} \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E}] \tag{24}$$

$$= \mu \cdot [\int_{\Upsilon} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E} - \int_{S} \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E}] \tag{25}$$

$$= \mu \cdot [\Pr(Z(\mathcal{X}) = 1) - \Pr(\mathcal{X} \in S)] \tag{26}$$

$$\geq 0. \tag{27}$$

Equation 23 is derived from 22 due to the fact that $\Pr(\mathcal{Y} = \mathcal{E})/\Pr(\mathcal{X} = \mathcal{E}) \leq \mu, \forall \mathcal{E} \in S$, $\Pr(\mathcal{Y} = \mathcal{E})/\Pr(\mathcal{X} = \mathcal{E}) \geq \mu, \forall \mathcal{E} \in S^c$, and $1 - Z(1|\mathcal{E}) \geq 0$. Similarly, we can establish the proof for part (2), but we have omitted the detailed steps for the sake of conciseness. □

For simplicity, we use $n_i$ and $n'_i$ to denote the number of basic elements (e.g., pixels) in $\mathbf{m}_i$ and $\mathbf{m}'_i$ respectively, i.e., $n_i = |\mathbf{m}_i|$ and $n'_i = |\mathbf{m}'_i|$. Then, we have the following probability mass function:

$$\Pr(\mathcal{X} = \mathcal{E}) = \begin{cases} \frac{1}{\prod_{i=1}^{T} \binom{n_i}{k_i}}, & \text{if } \mathcal{E} \in \tilde{A} \cup \tilde{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

$$\Pr(\mathcal{Y} = \mathcal{E}) = \begin{cases} \frac{1}{\prod_{i=1}^{T} \binom{n'_i}{k_i}}, & \text{if } \mathcal{E} \in \tilde{B} \cup \tilde{C}, \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

Recall that we have $e_i = |\mathbf{m}_i' \cap \mathbf{m}_i|$ for $i = 1, 2, \ldots, T$, so the probability of $\mathcal{X}$ and $\mathcal{Y}$ in $\tilde{A}$, $\tilde{B}$ and $\tilde{C}$ can be computed as follows:

$$\Pr(\mathcal{X} \in \tilde{A}) = 1 - \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i}{k_i}}, \Pr(\mathcal{X} \in \tilde{B}) = \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i}{k_i}}, \Pr(\mathcal{X} \in \tilde{C}) = 0; \tag{30}$$

$$\Pr(\mathcal{Y} \in \tilde{A}) = 0, \Pr(\mathcal{Y} \in \tilde{B}) = \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}}, \Pr(\mathcal{Y} \in \tilde{C}) = 1 - \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}}. \tag{31}$$

We first define $\delta_l = \underline{\Pr}(g(\mathcal{X}) = A) - \frac{\lfloor \underline{\Pr}(g(\mathcal{X})=A) \prod_{i=1}^{T} \binom{n_i}{k_i} \rfloor}{\prod_{i=1}^{T} \binom{n_i}{k_i}}$ to help rounding $\underline{\Pr}(g(\mathcal{X}) = A)$. Then we can construct a set $S = \tilde{A} + \tilde{B}'$, where $\tilde{B}' \subseteq \tilde{B}$ and $\Pr(\mathcal{X} \in \tilde{B}') = \underline{\Pr}(g(\mathcal{X}) = A) - \delta_l - \Pr(\mathcal{X} \in \tilde{A})$. We can assume $\underline{\Pr}(g(\mathcal{X}) = A) > \Pr(\mathcal{X} \in \tilde{A})$ because otherwise $\Pr(g(\mathcal{Y}) = A)$ is bounded by 0. Then we have $\Pr(g(\mathcal{X}) = A) \geq \Pr(\mathcal{X} \in S)$. So we have the following lower bound on $\Pr(g(\mathcal{Y}) = A)$:

$$\begin{aligned}
&\Pr(g(\mathcal{Y}) = A) && (32)\\
\geq &\Pr(\mathcal{Y} \in S) && (33)\\
\geq &\Pr(\mathcal{Y} \in \tilde{B}') && (34)\\
\geq &\Pr(\mathcal{X} \in \tilde{B}') \frac{\Pr(\mathcal{Y} \in \tilde{B}')}{\Pr(\mathcal{X} \in \tilde{B}')} && (35)\\
\geq &\frac{\prod_{i=1}^{T} \binom{n_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}} (\underline{\Pr}(g(\mathcal{X}) = A) - \delta_l - 1 + \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i}{k_i}}) && (36)
\end{aligned}$$

Similarly we define $\delta_u = \frac{\lceil \overline{\Pr}(g(\mathcal{X})=B) \prod_{i=1}^{T} \binom{n_i}{k_i} \rceil}{\prod_{i=1}^{T} \binom{n_i}{k_i}} - \overline{\Pr}(g(\mathcal{X}) = B)$, so we can construct a set $S = \tilde{B}' + \tilde{C}$, where $\tilde{B}' \subseteq \tilde{B}$ and $\Pr(\mathcal{X} \in \tilde{B}') = \overline{\Pr}(g(\mathcal{X}) = B) + \delta_u - \Pr(\mathcal{X} \in \tilde{C})$. Then we have $\Pr(g(\mathcal{X}) = B) \leq \Pr(\mathcal{X} \in S)$. So we have the following upper bound on $\Pr(g(\mathcal{Y}) = B)$:

$$\begin{aligned}
&\Pr(g(\mathcal{Y}) = B) && (37)\\
\leq &\Pr(\mathcal{Y} \in S) && (38)\\
\leq &\Pr(\mathcal{Y} \in \tilde{B}') + \Pr(\mathcal{Y} \in \tilde{C}) && (39)\\
\leq &\Pr(\mathcal{X} \in \tilde{B}') \frac{\Pr(\mathcal{Y} \in \tilde{B}')}{\Pr(\mathcal{X} \in \tilde{B}')} + \Pr(\mathcal{Y} \in \tilde{C}) && (40)\\
\leq &\frac{\prod_{i=1}^{T} \binom{n_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}} (\overline{\Pr}(g(\mathcal{X}) = B) + \delta_u) + \Pr(\mathcal{Y} \in \tilde{C}) && (41)\\
\leq &\frac{\prod_{i=1}^{T} \binom{n_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}} (\overline{\Pr}(g(\mathcal{X}) = B) + \delta_u) + 1 - \frac{\prod_{i=1}^{T} \binom{e_i}{k_i}}{\prod_{i=1}^{T} \binom{n_i'}{k_i}} && (42)
\end{aligned}$$

To certify a test sample, we just need to enforce $\Pr(g(\mathcal{Y}) = A) > \Pr(g(\mathcal{Y}) = B)$. So we get Theorem 1.

## B. Details About the Datasets

We use two benchmark datasets for evaluation.

- **RAVDESS.** We use RAVDESS dataset [33] for the multi-modal emotion recognition task. This dataset contains video recordings of 24 participants, each speaking with a variety of emotions. The goal is to classify these emotions into one of seven categories: calm, happy, sad, angry, fearful, surprise, and disgust. For each participant, there are 60 distinct video sequences. For data pre-processing, we follow previous work [2, 8] and crop or zero-pad these videos to 3.6 seconds, which

is the average video length. After pre-processing, each data sample contains 108 image frames and 79380 audio frames. We assume that the attacker can arbitrarily modify $r_1$ image frames (from 108 image frames of visual input) and $r_2$ audio frames (from 79380 audio frames of audio input). We divide the data into training, validation and test sets ensuring that the identities of actors are not repeated across sets. Particularly, we used four actors for testing, four for validation, and the remaining 16 for training.

- **KITTI Road.** For the multi-modal road segmentation task, we use KITTI Road Dataset [1], which contains 289 training and 290 test samples across three distinct road scene categories. Notably, the initial release [1] lacks ground-truth labels for its test samples. As a result, we divided the original training dataset into 231 data samples (80% of the data samples) for training and 58 data samples (20% of the data samples) for testing. Each data sample consists of a RGB image, a depth image, and the ground truth segmentation. We assume that the attacker can arbitrarily modify $r_1$ pixels from the RGB image and $r_2$ pixels from the depth image for each testing input.

## C. Special Cases in Multi-modal Segmentation

For segmentation tasks, the multi-modal model outputs the segmentation result for one of the input modalities $\mathbf{m}_o$ with $n_o$ basic elements, which can be pixels or 3-D points. Then the output contains $n_o$ labels. Previously, we consider the case where the attacker perform modification attacks to $\mathbf{m}_o$, where we have $\mathbf{m}_o = n_o = n'_o = |\mathbf{m}'_o|$. However, deletion and addition attacks on $\mathbf{m}_o$ are also possible if $\mathbf{m}_o$ represents a point cloud. If that is the case, the process of deriving Certified Pixel Accuracy, Certified F-score and Certified IoU can be different.

First, we think of the multi-modal segmentation model before the attack (denoted by $G$) as composed of multiple classifiers denoted by $G_1, G_2, \ldots, G_{n_o}$. Each classifier $G_j$ predicts a label $G_j(\mathbf{M})$ for $m_o^j$ (the $j$th basic element of $\mathbf{m}_o$). The ground truth $y$ also includes $n_o$ labels, denoted by $y_1, y_2, \ldots, y_{n_o}$. We use $G_j(\mathbf{M})$ to denote the predicted label for $m_o^j$ before the attack and use $G_j(\mathbf{M}')$ to denote the predicted label for $m_o^j$ after the attack. We say a basic element (e.g., a pixel) $m_o^j$ is *certifiably stable* if

$$G_j(\mathbf{M}) = G_j(\mathbf{M}'), \forall \mathbf{M}' \in \mathcal{S}(\mathbf{M}, \mathbf{R}), \text{and } m_o^j \in \mathbf{m}_o \cap \mathbf{m}'_o, \tag{43}$$

which means $j$th basic element of $\mathbf{m}_o$ is also in $\mathbf{m}'_o$ and the predicted label for it is unchanged by the attack. If it also holds that $G_j(\mathbf{M}) = y_j$, then we term $m_o^j$ as *certifiably robust*.

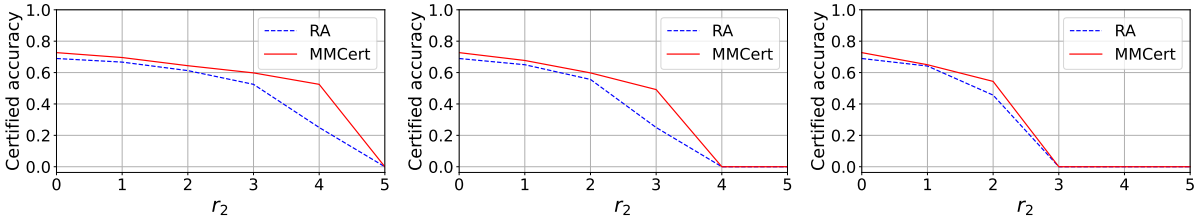
Then we derive Certified Pixel Accuracy (or F-score or IoU) for deletion and addition attacks on $\mathbf{m}_o$. We use $j \in [n_o]$ to denote the index of a basic element of the input modality $\mathbf{m}_o$. For each label, we define:

$$TP = |\{j : (G_j(\mathbf{M}) = y_j = 1) \wedge IsStable(\mathbf{M}, j)\}|,$$
$$TN = |\{j : (G_j(\mathbf{M}) = y_j = 0) \wedge IsStable(\mathbf{M}, j)\}|,$$
$$FP = |\{j : G_j(\mathbf{M}) = 1\}| - TP, \text{and}$$
$$FN = |\{j : G_j(\mathbf{M}) = 0\}| - TN,$$

where 1 indicates that this basic element has been identified as belonging to this label, while label 0 signifies the opposite. $IsStable(\mathbf{M}, j)$ is true if and only if the $j$th basic element of $\mathbf{m}_o$ is certifiably stable as defined above. We use $r_o$ denote the added (or deleted) basic elements for $\mathbf{m}_o$. Then for addition attacks to $\mathbf{m}_o$, the worst case is that all added basic elements are not certifiably robust, so we have Certified Pixel Accuracy $= \frac{TP+TN}{TP+TN+FP+FN+r_o}$, Certified F-score $= \frac{2TP^2}{2TP^2+TP(FP+FN+r_o)}$ , and Certified IoU $= \frac{TP}{TP+FP+FN+r_o}$. And for deletion attacks to $\mathbf{m}_o$, the worst case is that all deleted basic elements are certifiably robust, so we have Certified Pixel Accuracy $= \frac{TP+TN-r_o}{TP+TN+FP+FN-r_o}$, Certified F-score $= \frac{2(TP-r_o)^2}{2(TP-r_o)^2+(TP-r_o)(FP+FN)}$ , and Certified IoU $= \frac{TP-r_o}{TP+FP+FN-r_o}$. To obtain the final metrics, we compute the average of these values across all test samples and all labels.

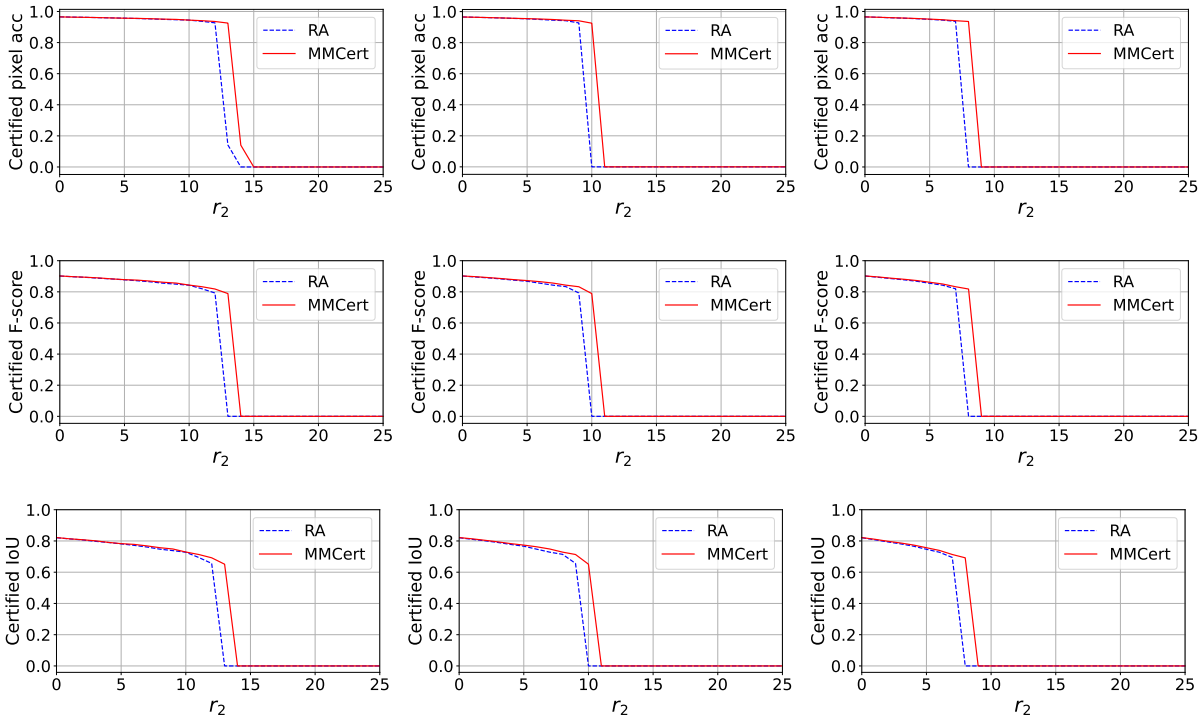
## D. Experiment Results for the $r_1 > r_2$ Case

Here, we compare our method with randomized ablation for the case $r_1 > r_2$. For KITTI Road dataset, we set $k_1$ to 4,000 and $k_2$ to 6,000 for our MMCert and set $k$ to 10,000 for randomized ablation. For RAVEDESS, we let $k_1$ = 5 and $k_2$ = 1,000 for our MMCert and let $k$ = 3,000 for randomized ablation. The results of these experiments are illustrated in Figures 5 and 6, corresponding to RAVNESS and KITTI Road datasets, respectively. Our findings reveal that our method consistently



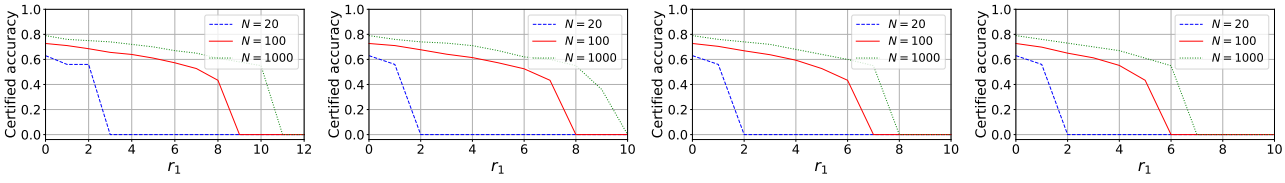

(a) $r_1 = 2r_2$ (b) $r_1 = 3r_2$ (c) $r_1 = 4r_2$

Figure 5. Compare our MMCert with randomized ablation on RAVDESS Dataset.




(a) $r_1 = 2r_2$ (b) $r_1 = 3r_2$ (c) $r_1 = 4r_2$

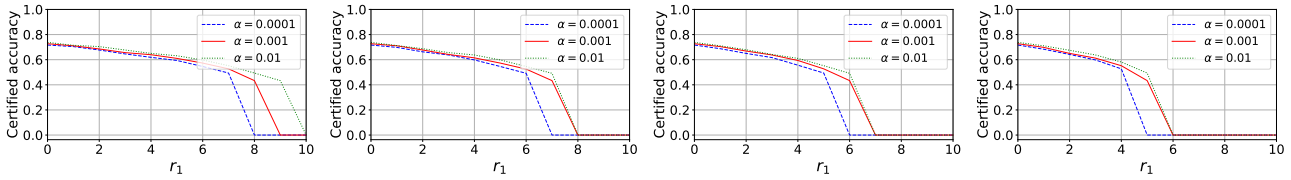
Figure 6. Compare our MMCert with randomized ablation on KITTI Road Dataset. Certified Pixel Accuracy (first row), Certified F-score (second row) and Certified IoU (third row) are considered.




(a) $r_1 = r_2$ (b) $r_1 = 2r_2$ (c) $r_1 = 3r_2$ (d) $r_1 = 4r_2$

Figure 7. Impact of $N$ on RAVDESS dataset.

surpasses randomized ablation across all $r_1$-$r_2$ ratios for both datasets. This can be attributed to the fact that randomized ablation is essentially a special case of our MMCert. Consequently, we can identify a combination of $k_1$ and $k_2$ that yields equal or better results than randomized ablation. Furthermore, our MMCert is more stable than randomized ablation during both training and testing phases because our method's sub-sampled input space is smaller than that of randomized ablation.




(a) $r_1 = r_2$ (b) $r_1 = 2r_2$ (c) $r_1 = 3r_2$ (d) $r_1 = 4r_2$

Figure 8. Impact of $\alpha$ on RAVDESS dataset.

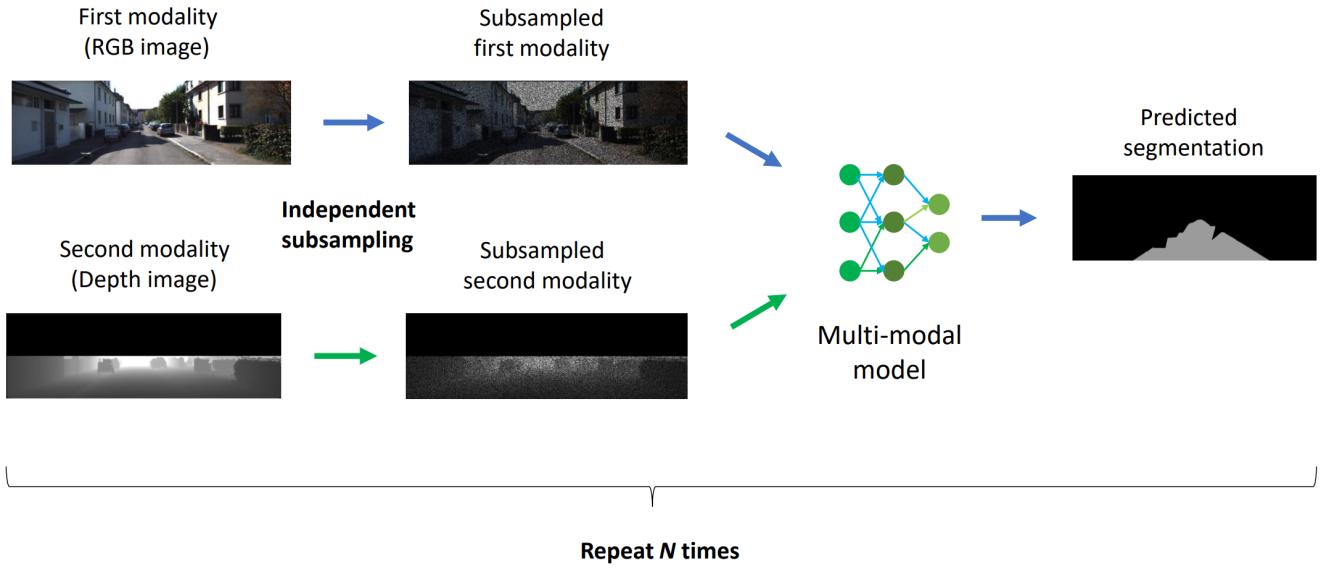



Figure 9. Illustration of independent sub-sampling on KITTI Road dataset. Our method repeatedly generate predictions for subsampled multi-modal inputs. These predictions are then aggregated to get the final prediction.

## E. Impact of $N$ and $\alpha$

We study the impact of $N$ and $\alpha$ on RAVDESS dataset. Figure 7 in Appendix shows the impact of $N$. We discover that the certified accuracy improves with an increase in $N$. This enhancement occurs because a larger $N$ yields tighter lower or upper bounds for the label probability, given a constant confidence level $\alpha$. However, the computational cost also grows linearly with respect to $N$, reflecting a trade off between computational cost and certification performance. Figure 8 in Appendix shows the impact of $\alpha$. We observe that MMCert achieves better performance as $\alpha$ increases. This shows the trade off between the confidence of the certification and the certification performance.