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 $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_T)$ and attacked input pair $(\mathbf{m}'_1, \mathbf{m}'_2, \dots, \mathbf{m}'_T)$, 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 Υ to denote the joint space between $\mathcal{X}$ and $\mathcal{Y}$. We use $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_T)$ to denote a variable in the space Υ.

We divide the space Υ into the following subspace:

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

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

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

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:

$$
\begin{aligned}
&\Pr(Z(\mathcal{Y}) = 1) - \Pr(\mathcal{Y} \in S) && (19)\\
=&\int_{\Upsilon} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} - \int_S \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} && (20)\\
=&\int_{S^c} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} + \int_S Z(1|\mathcal{E}) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} - \int_S \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} && (21)\\
=&\int_{S^c} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} - \int_S (1 - Z(1|\mathcal{E})) \cdot \Pr(\mathcal{Y} = \mathcal{E}) d\mathcal{E} && (22)\\
\geq&\mu \cdot [\int_{S^c} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E} - \int_S (1 - Z(1|\mathcal{E})) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E}] && (23)\\
=&\mu \cdot [\int_{S^c} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E} + \int_S Z(1|\mathcal{E}) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E} - \int_S \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E}] && (24)\\
=&\mu \cdot [\int_{\Upsilon} Z(1|\mathcal{E}) \cdot \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E} - \int_S \Pr(\mathcal{X} = \mathcal{E}) d\mathcal{E}] && (25)\\
=&\mu \cdot [\Pr(Z(\mathcal{X}) = 1) - \Pr(\mathcal{X} \in S)] && (26)\\
\geq&0. && (27)
\end{aligned}
$$

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 $\Pr(g(\mathcal{Y}) = A)$:

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

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 $\Pr(g(\mathcal{Y}) = B)$:

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

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 $G_1, G_2, \dots, G_{n_o}$. Each classifier G_j predicts a label $G_j(\mathbf{M})$ for m_o^j (the jth basic element of $\mathbf{m}_o$). The ground truth y also includes n_o labels, denoted by $y_1, y_2, \dots, y_{n_o}$. 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 jth 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:

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

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

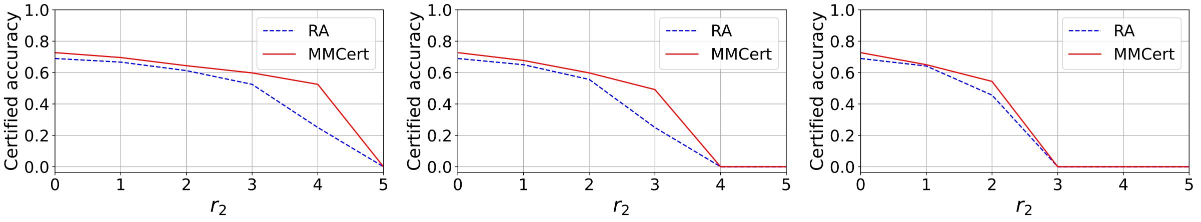
$$FP = |\{j : G_j(\mathbf{M}) = 1\}| - TP, \text{and}$$

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

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 jth 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.

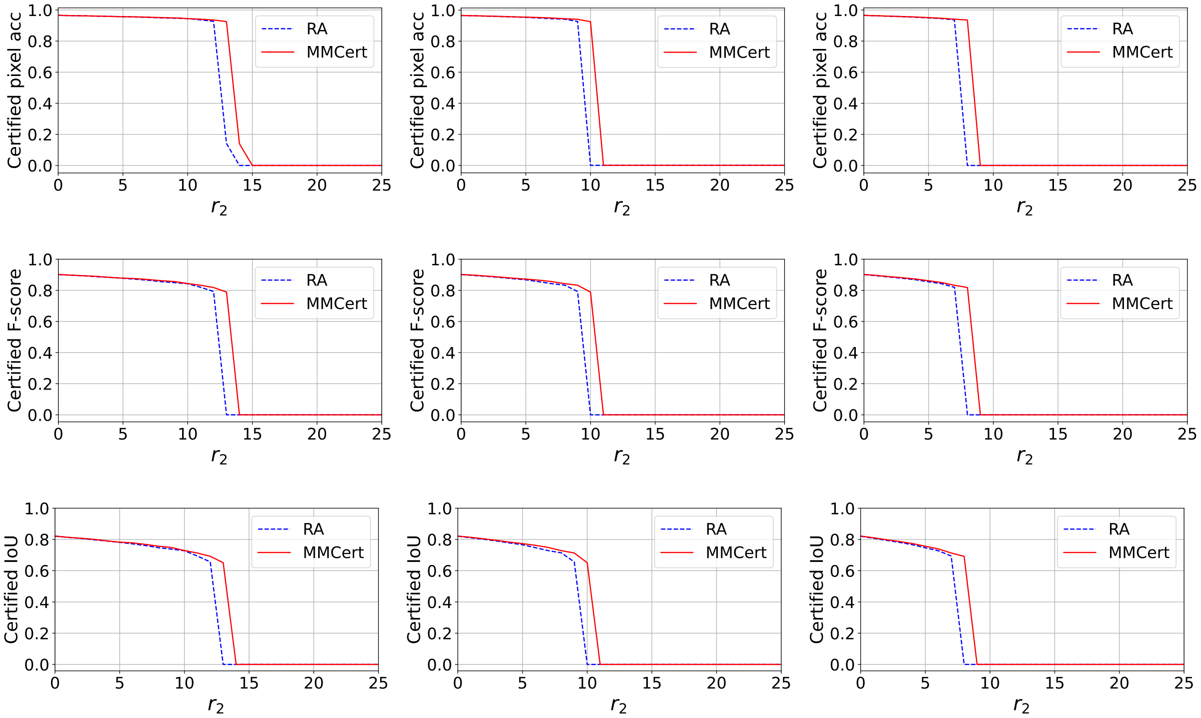
D. Experiment Results for the $r_1 > r_2$ Case

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



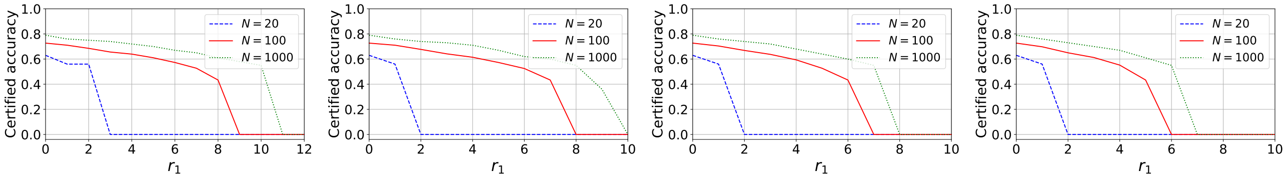
(a) $r_1 = 2r_2$ (b) $r_1 = 3r_2$ (c) $r_1 = 4r_2$

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



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

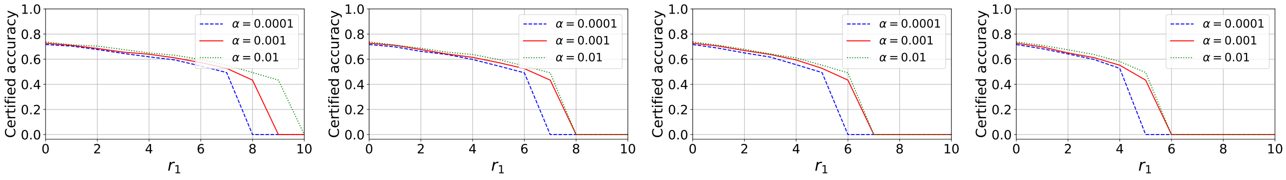
Figure 6. Compare our MMCert with randomized ablation on KITTI Road Dataset. Certified Pixel Accuracy (first row), Certified F-score (second row) and Certified IoU (third row) are considered.



(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 α on RAVDESS dataset.

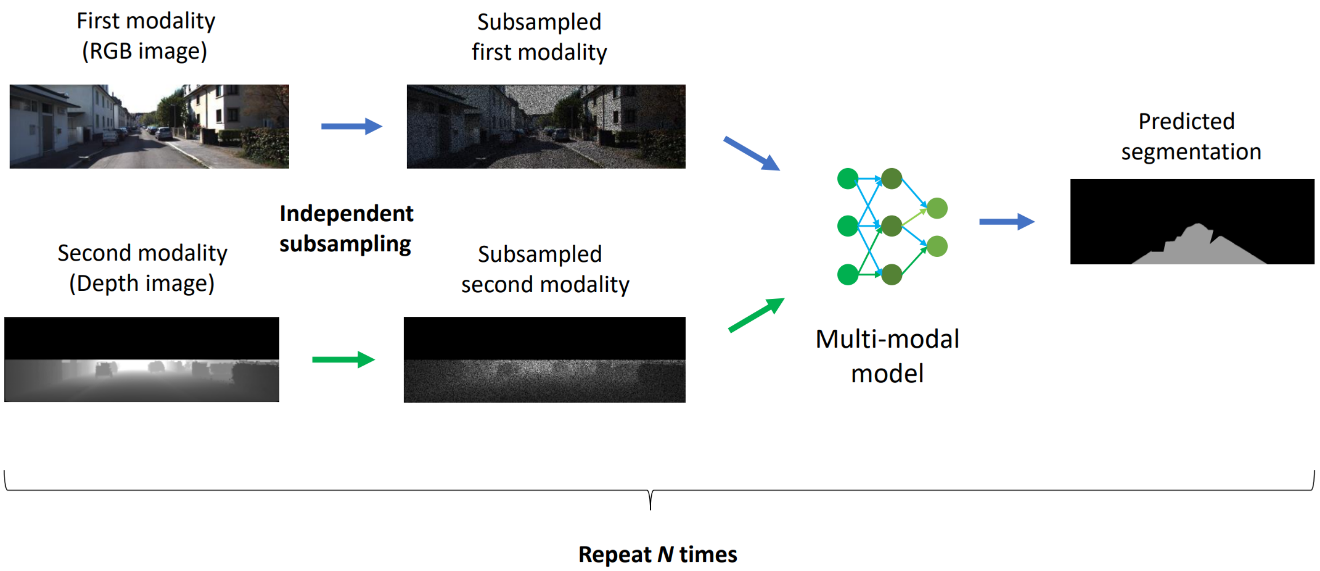


Figure 9. Illustration of independent sub-sampling on KITTI Road dataset. Our method repeatedly generate predictions for subsampled multi-modal inputs. These predictions are then aggregated to get the final prediction.

E. Impact of N and α

We study the impact of N and α 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 α. 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 α. We observe that MMCert achieves better performance as α increases. This shows the trade off between the confidence of the certification and the certification performance.