FUSIONBREAK: ATTACKING MULTIMODAL FUSION WHILE PRESERVING UNIMODAL DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models combine information from multiple modalities, but correct unimodal predictions do not necessarily guarantee a correct fused prediction. Conventional adversarial evaluations typically focus on whether the final multimodal prediction becomes incorrect without distinguishing errors that emerge at the fused output from those already present in individual modalities. We study a stricter condition in which the unimodal classifiers remain correct while the fused model becomes incorrect. We introduce FusionBreak, an adversarial attack that preserves unimodal decisions while causing the fused model to misclassify, and define the Fusion-Only Success Rate (FOSR) to measure this behavior. Experiments across two datasets and four fusion schemes show that fusion-only failures occur across different fusion designs. In contrast, standard PGD can achieve a high fused attack success rate with 0% FOSR, showing that conventional attack success does not distinguish whether the fused model fails while the unimodal predictions remain correct. Ablation experiments further demonstrate the importance of preserving unimodal decisions. We also evaluate a detector-aware variant of FusionBreak that adds detector-evasion terms to the attack objective. This reduces detection by several generic adversarial detectors while largely preserving the effectiveness of the attack.

1 INTRODUCTION

Multimodal models combine information from different sources to improve prediction accuracy. For example, an audio-visual recognition system can use both speech and facial motion when making a decision. Because the modalities provide complementary information, their combination is often expected to make the final system more reliable. However, multimodal models can still be vulnerable when one or more inputs are noisy, unreliable, or adversarially manipulated (Yu et al., 2020; Yang et al., 2021; Tian & Xu, 2021). Therefore, several fusion methods have been proposed to estimate modality reliability and reduce the influence of unreliable inputs (Zhang et al., 2023; Gao et al., 2024).

Most adversarial evaluations of multimodal systems measure success only at the final fused output (Evtimov et al., 2020; Vishwamitra et al., 2021; Zhang et al., 2022; Lu et al., 2023). If the fused model becomes incorrect, the attack is counted as successful regardless of what happens to the individual unimodal models. This creates an important ambiguity. The attack may have already caused the audio classifier, the visual classifier, or both to make an incorrect prediction, in which case the fused error may propagate from an incorrect unimodal prediction rather than reveal a vulnerability in the fusion process itself. As a result, conventional attack success rates cannot distinguish failures that occur despite correct unimodal predictions from errors that originate within the individual modalities. This distinction matters because we need to know whether an attacker can manipulate otherwise correct modality-specific information into an incorrect final decision when evaluating fusion robustness.

We are interested in a different situation: both unimodal classifiers still make the correct prediction, but the fused model becomes incorrect. This distinction can matter in practical multimodal systems. For example, consider two organizations that independently process different modalities. One organization operates an audio classifier, while another operates a visual classifier. Each organization may verify its own prediction before sharing information with a joint system. If both local

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

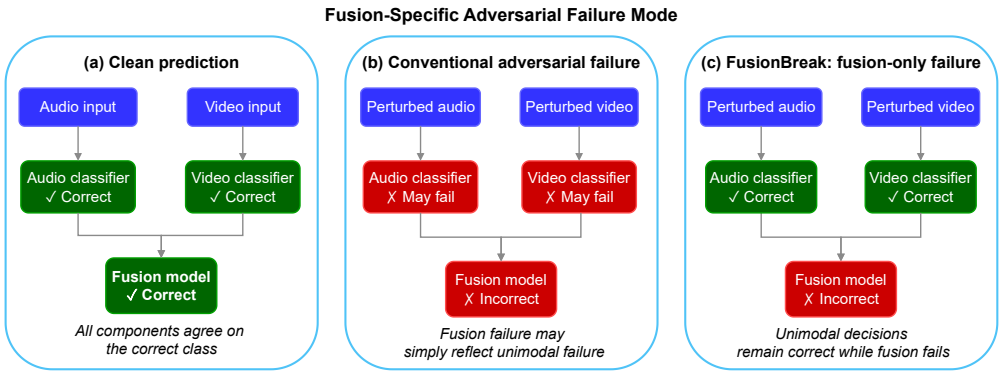

Figure 1: Comparison of clean prediction, conventional adversarial failure, and the fusion-only failure targeted by FusionBreak. In a fusion-only failure, the audio and visual classifiers remain correct while the fused model outputs an incorrect prediction.

classifiers remain correct, neither organization observes an incorrect prediction in its own modality. However, after their representations are combined, the shared fused model may still make an incorrect decision.

This example shows why preserving unimodal correctness is a useful attack constraint. It separates an incorrect prediction produced after fusion from an incorrect prediction that was already present in one of its components. It can also expose a weakness that would be missed by a monitoring strategy that checks only whether the individual classifiers continue to output the expected labels. We do not assume that modality-specific adversarial detectors would necessarily fail, but unimodal prediction correctness alone is not sufficient to guarantee a correct fused prediction. We call this condition a *fusion-only failure*, as illustrated in Figure 1.

The main novelty of our work is to make *unimodal correctness an explicit requirement of the adversarial attack*. We ask whether a fused classifier can be fooled *while all of its unimodal classifiers remain correct*. This gives a direct way to distinguish fusion-only failures from ordinary adversarial failures of the individual modalities. We propose *FusionBreak*, which attacks the fused model under this constraint. We evaluate attacks on samples for which the clean audio, visual, and fused models are all correct; we refer to these as *eligible samples*. We introduce the *Fusion-Only Success Rate* (FOSR), defined as the fraction of eligible samples for which the fused model becomes incorrect while both unimodal classifiers remain correct after the attack. We also report fusion attack success rate (FASR), which measures fused misclassification on the same eligible samples regardless of the unimodal predictions.

Our experiments demonstrate an important difference between conventional attack success and fusion-only attack success. FusionBreak achieves a high FOSR on both RAVDESS (Livingstone & Russo, 2018) and CREMA-D (Cao et al., 2014), while standard PGD achieves complete fused attack success but no fusion-only success.

We further evaluate a detector-aware variant of FusionBreak by adding detector-evasion terms to the attack objective while preserving the same fusion-only objective. This allows us to evaluate whether fusion-only adversarial examples remain effective when the attack is optimized against adversarial detectors.

2 Background

This section reviews prior works related to the problem studied in this paper. We first discuss the reliability of multimodal fusion and how different fusion mechanisms respond to unreliable modalities. We then review adversarial attacks on multimodal systems, with particular attention to audio-visual settings, followed by adversarial detection methods relevant to our evaluation.

2.1 Reliability of Multimodal Fusion

Multimodal learning combines information from different sources to improve prediction performance (Yu et al., 2020; Zhang et al., 2026b). Fusion can be performed through concatenation, averaging, attention, learned gating, and other mechanisms (Evtimov et al., 2020; Zhang et al., 2026b; Yang et al., 2021). However, higher clean performance does not always imply better robustness. A noisy or corrupted modality can strongly influence the final prediction and cause the fused model to fail (Yu et al., 2020; Yang et al., 2021).

Robustness can also decrease when modalities differ in quality, are missing, or contribute unevenly to the prediction (Zhang et al., 2026b). Differences between the modalities available during training and deployment can create similar problems (Mckinzie et al., 2023). Prior work has also shown that models with similar clean accuracy may use multimodal information differently and therefore respond differently to corruption (Kim & Ghosh, 2019). Modalities can also have different gradient sensitivities and different effects on overall robustness (Zhang et al., 2026a).

The fused prediction may also disagree with otherwise useful unimodal predictions. Prior work has linked such disagreement to ambiguous or conflicting information across modalities (Srikanth et al., 2025). This indicates that disagreement can arise after information from the modalities is combined, even when the individual modalities remain informative. Such behavior motivates evaluating the fused prediction together with the unimodal predictions rather than evaluating each modality independently.

2.2 Adversarial Attacks on Multimodal Systems

Adversarial examples are small input modifications designed to cause an incorrect model prediction. Adversarial attacks have been studied across a wide range of objectives, including attacks on model utility, privacy, and explainability (Vashagh et al., 2026; Meymani et al., 2026). Projected Gradient Descent (PGD) is widely used for robustness evaluation because it iteratively increases the classification loss while keeping the perturbation within a fixed budget (Madry et al., 2018). In multimodal systems, an attacker may modify one modality or several modalities simultaneously.

Previous work has shown that attacking a single sensor can reduce the performance of a fused model even when other inputs remain unchanged. This has been studied in camera-LiDAR detection (Park et al., 2020; Tu et al., 2022), RGB-thermal segmentation (Yu et al., 2020), and 3D object detection (Long et al., 2025). Multimodal attacks have also been studied under partial-knowledge settings (Evtimov et al., 2020).

Other attacks target multimodal representations or interactions more directly. MUROAN searches for image regions or text tokens whose removal changes the fused representation and final prediction (Vishwamitra et al., 2021). Attacks on vision-language models have similarly used joint image-text perturbations, cross-modal alignment, and attention information to fool multimodal predictions (Zhang et al., 2022; Lu et al., 2023; Wang et al., 2024a; Guan et al., 2024; Bagdasaryan et al., 2024).

These studies show that attackers can target both individual modalities and their interactions in multimodal models. However, attack success is usually defined using the final multimodal prediction. Existing attacks generally do not require the unimodal classifiers to remain correct after the attack. As a result, a successful attack may reflect a failure of one of the individual modalities rather than a failure specific to fusion.

2.3 Audio-Visual Adversarial Robustness

Audio-visual systems are especially relevant to our setting because audio and video provide different but related information. Tian & Xu (2021) studied audio-only, video-only, and joint attacks and showed that attacking one modality can reduce the benefit provided by the other. They also showed that robustness depends on the fusion method.

Large-scale evaluations have similarly found that joint audio-visual attacks can be stronger than single-modality attacks and that modality dominance affects robustness (Li et al., 2022). Recently, Zhang et al. (2025) studied attacks based on temporal invariance and modality misalignment, using temporal transformations and reduced audio-visual similarity to weaken multimodal interaction.

These studies establish that audio-visual systems can fail because of attacks on individual modalities or their interaction. Our work focuses on a different condition: whether the fused model can be made incorrect while the audio and visual classifiers themselves remain correct.

2.4 Multimodal Defenses

Several defenses have been proposed for multimodal systems. Contextual fusion allows different representations to compensate for one another under corruption (Akumalla et al., 2020), while other approaches estimate modality reliability and reduce the influence of unreliable inputs (Yang et al., 2021). MMCert provides certified robustness against bounded sparse modifications (Wang et al., 2024b), and VARMAT adapts training according to the vulnerability of different modalities (Zhang et al., 2026a).

Adversarial detection provides another form of defense. Maximum softmax probability (MSP) uses model confidence as an abnormality signal (Hendrycks & Gimpel, 2017). Feature squeezing compares predictions before and after simple input transformations (Xu et al., 2018), while Local Intrinsic Dimensionality (LID) examines the local structure of hidden representations (Ma et al., 2018). Mahalanobis-based methods instead measure the distance between a test representation and clean class-conditional representations (Lee et al., 2018).

Multimodal systems also provide an additional signal: consistency between unimodal and fused predictions. For example, FlowGuard uses differences between unimodal and multimodal outputs to identify unsafe cross-modal interactions (Yeon et al., 2026), while Liu et al. (2026) consider confidence differences between fused and unimodal models. These works use unimodal–multimodal disagreement as a signal for detecting unreliable multimodal predictions. In contrast, our work does not propose a defense based on such disagreement. Instead, we study whether an attacker can deliberately create this specific failure condition: the fused classifier becomes incorrect while the unimodal classifiers remain correct. We then use existing detection methods, including consistency-based signals, to evaluate how detectable these fusion-only adversarial examples are.

3 Proposed Method

We first define the problem setting and threat model, followed by the FusionBreak objective, its detector-aware extension, and the optimization and candidate-selection procedure.

3.1 Problem Setting

Throughout this section, the subscripts a, v, and F refer to audio, visual, and fused quantities, respectively. We consider an audio-visual classification problem with C classes, indexed by $c \in \{1, \dots, C\}$. For a sample (x_a, x_v, y), x_a and x_v denote the processed audio and visual inputs, respectively, and $y \in \{1, \dots, C\}$ is the ground-truth label.

Let $h_a(x_a), h_v(x_v) \in \mathbb{R}^C$ denote the unimodal logits and $h_F(x_a, x_v) \in \mathbb{R}^C$ the fused logits. The corresponding predicted-class functions are $f_a(x_a) = \arg\max_c [h_a(x_a)]_c$, $f_v(x_v) = \arg\max_c [h_v(x_v)]_c$, and $f_F(x_a, x_v) = \arg\max_c [h_F(x_a, x_v)]_c$. Let $(\tilde{x}_a, \tilde{x}_v)$ denote the final adversarial pair returned by the attack.

We evaluate attacks on the eligible set $\mathcal{E} = \{(x_a, x_v, y) : f_a(x_a) = y,\ f_v(x_v) = y,\ f_F(x_a, x_v) = y\}$. Let $|\mathcal{E}|$ denote its cardinality and $\mathbf{1}[\cdot]$ the indicator function. The fused attack success rate (FASR) is

$$\mathrm{FASR} = \frac{1}{|\mathcal{E}|} \sum_{(x_a, x_v, y) \in \mathcal{E}} \mathbf{1}[f_F(\tilde{x}_a, \tilde{x}_v) \neq y], \tag{1}$$

and the fusion-only success rate (FOSR) is

$$\mathrm{FOSR} = \frac{1}{|\mathcal{E}|} \sum_{(x_a, x_v, y) \in \mathcal{E}} \mathbf{1}\big[f_F(\tilde{x}_a, \tilde{x}_v) \neq y,\ f_a(\tilde{x}_a) = y,\ f_v(\tilde{x}_v) = y\big]. \tag{2}$$

Thus, FASR measures fused misclassification regardless of the unimodal predictions, while FOSR additionally requires both unimodal classifiers to remain correct.

We consider a white-box, untargeted attacker with access to the audio, visual, and fused models, including their parameters and gradients. Let $\epsilon_a > 0$ and $\epsilon_v > 0$ denote the audio and visual L_∞ perturbation budgets, respectively. The attacker can perturb both modalities within these budgets but cannot modify the models or the ground-truth label. For the detector-aware variant, the attacker also has access to the detectors in $\mathcal{D}_{\text{opt}}$, including their scores and decision thresholds.

Let d denote the unimodal representation dimension, with $z_a(x_a), z_v(x_v) \in \mathbb{R}^d$. We consider four fusion schemes: Concatenation, Interaction, Gated-I, and Gated-II. Concatenation uses $\phi_{\text{cat}} = [z_a, z_v]$, while Interaction augments the concatenated representation with $z_a \odot z_v$ and $|z_a - z_v|$. Gated-I computes $g = \sigma(W_g[z_a, z_v] + b_g)$. For RAVDESS, it uses $\phi_{\text{G1}} = g \odot z_a + (1_d - g) \odot z_v$, while for CREMA-D it uses $\phi_{\text{G1}} = [g \odot z_a, (1_d - g) \odot z_v]$. Gated-II applies an additional constrained gating formulation: for RAVDESS, a learned $2d$-dimensional sigmoid gate is applied to $[z_a, z_v]$, while for CREMA-D the Gated-I representation uses $\bar{g} = \text{round}(4g)/4$ in place of g.

3.2 FusionBreak Objective

For logits $h \in \mathbb{R}^C$, define the true-class margin as $m(h, y) = h_y - \max_{c \neq y} h_c$, the softmax probabilities as $p(h) = \text{softmax}(h)$, and the cross-entropy (CE) loss as $\text{CE}(h, y) = -\log p(h)_y$.

For a generic perturbed candidate (x'_a, x'_v), define $m_a = m(h_a(x'_a), y)$, $m_v = m(h_v(x'_v), y)$, and $m_F = m(h_F(x'_a, x'_v), y)$. We use $\mathcal{L}^F_{\text{CE}} = \text{CE}(h_F(x'_a, x'_v), y)$, $\mathcal{L}^F_{\text{mar}} = -m_F$, and $\mathcal{L}_{\text{pres}} = [\rho - m_a]_+ + [\rho - m_v]_+$, where $[u]_+ = \max(u, 0)$ and $\rho > 0$ is the desired unimodal margin.

To keep perturbed unimodal outputs close to their clean counterparts, let $\bar{h}_a = h_a(x_a)$ and $\bar{h}_v = h_v(x_v)$. For probability vectors $\boldsymbol{\pi}, \boldsymbol{\kappa} \in [0, 1]^C$, define $D_{\text{KL}}(\boldsymbol{\pi} \| \boldsymbol{\kappa}) = \sum_{c=1}^{C} \pi_c \log(\pi_c / \kappa_c)$. The output-consistency loss is $\mathcal{L}_{\text{KL}} = D_{\text{KL}}(p(\bar{h}_a) \| p(h_a(x'_a))) + D_{\text{KL}}(p(\bar{h}_v) \| p(h_v(x'_v)))$. The complete FusionBreak objective is

$$\mathcal{J}_{\text{FB}} = \mathcal{L}^F_{\text{CE}} + \lambda_F \mathcal{L}^F_{\text{mar}} - \lambda_M \mathcal{L}_{\text{pres}} - \lambda_K \mathcal{L}_{\text{KL}}, \tag{3}$$

where $\lambda_F, \lambda_M, \lambda_K \geq 0$ control the fused-margin, unimodal-preservation, and output-consistency terms, respectively.

3.3 Detector-Aware Objective

For detector-aware optimization, let $\mathcal{D}_{\text{opt}}$ denote the set of detectors explicitly optimized against. For each $j \in \mathcal{D}_{\text{opt}}$, let $s_j(x'_a, x'_v)$ denote its score for the current perturbed candidate, τ_j its detection threshold, $q_j > 0$ the standard deviation of its clean-validation scores, and $\omega_j \geq 0$ its weight. Define the normalized threshold exceedance as $e_j = [(s_j(x'_a, x'_v) - \tau_j)/q_j]_+$ and the detector penalty as $\mathcal{L}_D = \sum_{j \in \mathcal{D}_{\text{opt}}} \omega_j e_j^2$. The detector-aware objective is

$$\mathcal{J}_{\text{FB+D}} = \mathcal{J}_{\text{FB}} - \lambda_D \mathcal{L}_D, \tag{4}$$

where $\lambda_D \geq 0$ controls the detector-evasion term. In our experiments, $\mathcal{D}_{\text{opt}}$ contains MSP, Mahalanobis distance, and feature squeezing. Feature squeezing uses a straight-through approximation during optimization. LID is evaluated but is not optimized against because its nearest-neighbor-based score is not directly suited to the gradient-based procedure.

3.4 Optimization and Candidate Selection

Let $B_a(x_a)$ and $B_v(x_v)$ denote the feasible perturbation sets satisfying $\|x'_a - x_a\|_\infty \leq \epsilon_a$ and $\|x'_v - x_v\|_\infty \leq \epsilon_v$, respectively. Perturbations are applied in the processed-input space. For CREMA-D, video projection additionally clips values to $[0, 1]$. For any feasible set B, $\Pi_B(\cdot)$ denotes projection onto that set.

We use $\mathcal{J} = \mathcal{J}_{\text{FB}}$ for FusionBreak and $\mathcal{J} = \mathcal{J}_{\text{FB+D}}$ for its detector-aware variant. Let R denote the number of restarts and $r \in \{1, \ldots, R\}$ the restart index. Let $\alpha_a, \alpha_v > 0$ denote the audio and visual

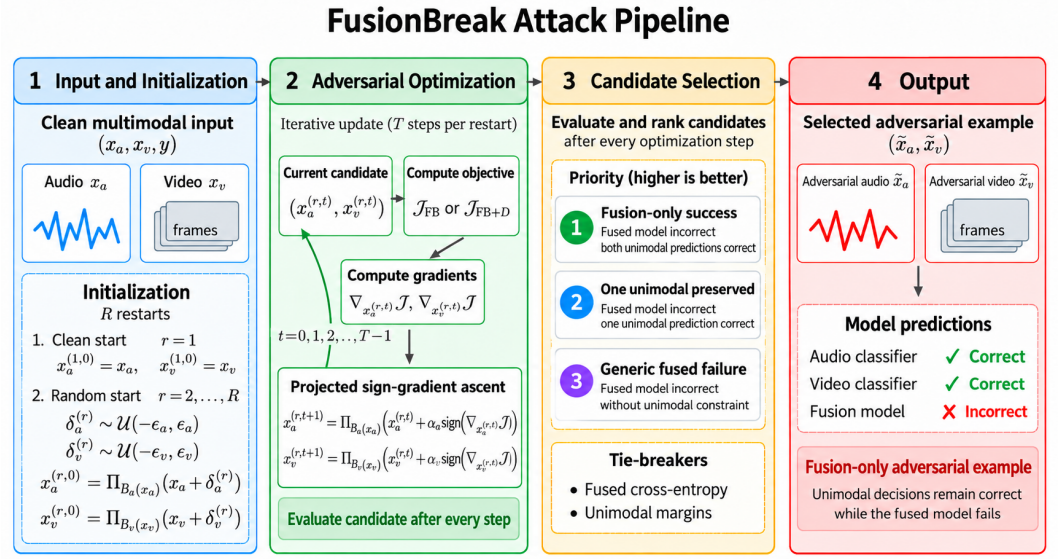

Figure 2: FusionBreak pipeline. Each restart performs projected sign-gradient updates and evaluates candidates after every step. Candidates are ranked by fusion-only success, unimodal preservation, and tie-breakers, and the highest-ranked pair is returned.

step sizes, T the number of optimization steps per restart, and $t \in \{0, \ldots, T-1\}$ the step index. We denote the candidate at step t of restart r by $(x_a^{(r,t)}, x_v^{(r,t)})$. Projected sign-gradient ascent gives

$$
\begin{aligned}
x_a^{(r,t+1)} &= \Pi_{\mathcal{B}_a(x_a)} \left(x_a^{(r,t)} + \alpha_a \, \text{sign} \left(\nabla_{x_a^{(r,t)}} \mathcal{J} \right) \right), \\
x_v^{(r,t+1)} &= \Pi_{\mathcal{B}_v(x_v)} \left(x_v^{(r,t)} + \alpha_v \, \text{sign} \left(\nabla_{x_v^{(r,t)}} \mathcal{J} \right) \right).
\end{aligned}
\tag{5}
$$

The first restart begins from the clean input, $x_a^{(1,0)} = x_a$ and $x_v^{(1,0)} = x_v$. For $r > 1$, random perturbations $\delta_a^{(r)} \sim \mathcal{U}(-\epsilon_a, \epsilon_a)$ and $\delta_v^{(r)} \sim \mathcal{U}(-\epsilon_v, \epsilon_v)$ are sampled element-wise, and the restart is initialized as $x_a^{(r,0)} = \Pi_{\mathcal{B}_a(x_a)}(x_a + \delta_a^{(r)})$ and $x_v^{(r,0)} = \Pi_{\mathcal{B}_v(x_v)}(x_v + \delta_v^{(r)})$.

Candidates are evaluated after every optimization step. A fusion-only success, satisfying $f_F(x'_a, x'_v) \neq y$, $f_a(x'_a) = y$, and $f_v(x'_v) = y$, receives the highest priority. Among the remaining candidates, a fused failure preserving one unimodal prediction is preferred over a fused failure preserving neither. Fused cross-entropy and unimodal margins are used as tie-breakers. After all R restarts, the highest-ranked candidate is returned as $(\tilde{x}_a, \tilde{x}_v)$. Figure 2 summarizes the complete FusionBreak procedure, including multi-restart initialization, projected sign-gradient optimization, candidate evaluation, and selection of the final fusion-only adversarial example.

4 Experimental Setup

We evaluate seeds 42, 43, and 44 across the four fusion schemes described in Section 3.1. Unless otherwise stated, results are reported as mean and standard deviation across the resulting 12 seed–scheme configurations.

We evaluate on the RAVDESS and CREMA-D audio-visual emotion-recognition datasets, which contain eight and six emotion classes, respectively. For each random seed, the data are divided into training, validation, and test sets, using 70/15/15% splits for RAVDESS and 70/10/20% splits for CREMA-D. RAVDESS uses paired speech audio and audio-video recordings, while CREMA-D uses matched audio and video utterances.

Table 1: Clean test accuracy and eligible-sample rate. Values are mean ± standard deviation across three seeds and four fusion schemes.

Dataset	Audio	Video	Fusion	Eligible
RAVDESS	47.22 ± 3.81	33.49 ± 5.55	63.12 ± 5.07	15.39 ± 1.91
CREMA-D	46.92 ± 3.05	45.62 ± 4.06	64.25 ± 4.93	20.16 ± 2.85

Audio is converted to a fixed-size time-frequency representation, while the video branch processes sampled frames and aggregates information across time. The audio encoder uses convolutional layers with 32, 64, and 128 channels, and both unimodal branches produce 128-dimensional representations before fusion. The unimodal classifiers are trained first, followed by the fusion models. Training uses cross-entropy loss and AdamW with a batch size of 64 and weight decay 10^{-4}, with early stopping based on validation accuracy. Dataset-specific preprocessing, augmentation, and training details are provided in Appendix A.

For the main attacks, we use $T = 80$ optimization steps and $R = 10$ restarts with perturbation budgets $\epsilon_a = 0.08$ and $\epsilon_v = 0.06$, and step sizes $\alpha_a = 0.004$ and $\alpha_v = 0.003$. The FusionBreak objective uses $\lambda_F = 0.5$, $\lambda_M = 5.0$, $\lambda_K = 0.25$, and $\rho = 0.05$; the detector-aware variant additionally uses $\lambda_D = 2.0$. Detector thresholds are calibrated on clean validation data to a 5% false-positive rate. Additional projection, attack, and detector implementation details are provided in Appendix A.

Standard multimodal PGD is the main baseline. It maximizes only the fused cross-entropy under the same perturbation limits, without unimodal preservation or decision-aware candidate selection.

5 Results

We evaluate FusionBreak, its detector-aware variant, standard PGD, controlled baselines, and single-modality attacks. All attacks use the full eligible test sets defined in Section 3.1. Across the 12 configurations, there are 399 attacked samples on RAVDESS and 3,598 on CREMA-D. Table 1 reports the corresponding eligible rates.

5.1 Fusion-Only Attack Performance

Table 2 reports the main attack results and controls. PGD reaches 100% FASR on both datasets but has 0% FOSR because the fused prediction changes while the unimodal predictions are usually also disrupted. Thus, fused misclassification alone does not indicate a fusion-only failure.

FusionBreak reaches $77.09 \pm 7.92\%$ FOSR on RAVDESS and $97.67 \pm 1.75\%$ on CREMA-D. Its detector-aware variant reaches $75.99 \pm 8.80\%$ and $97.64 \pm 1.91\%$, respectively. The dataset difference coincides with differences in unimodal preservation. FusionBreak preserves the video prediction in 86.72% of RAVDESS attacks and 97.73% of CREMA-D attacks. RAVDESS also has lower fused attack success and audio preservation, making simultaneous fused failure and unimodal preservation more difficult.

The attack works across all four fusion designs. Table 5 in Appendix B reports the results for each scheme. FOSR ranges from 74.38% to 80.83% on RAVDESS and from 97.13% to 98.44% on CREMA-D, so the overall result is not driven by a single fusion design.

The detector-aware objective has little effect on FOSR. Relative to FusionBreak, FOSR decreases by only 1.10 percentage points on RAVDESS and 0.03 percentage points on CREMA-D, indicating that detector awareness can be incorporated without materially reducing fusion-only attack success.

5.2 Controls and Detector-Aware Evaluation

The controls in Table 2 separate the effect of the FusionBreak objective from the search procedure. The CE-only baseline uses the same decision-aware search but removes the unimodal-preservation terms. Although it reaches 100% FASR on both datasets, its FOSR falls to $10.02 \pm 6.99\%$ on RAVDESS and $18.68 \pm 5.05\%$ on CREMA-D. Decision-aware search alone therefore does not explain the high FOSR of FusionBreak.

Table 2: Attack and control results. Values are mean ± standard deviation across three seeds and four fusion schemes. Audio and video preservation are the fractions of eligible samples for which the corresponding unimodal classifier remains correct after the attack. All values are percentages.

Dataset	Attack	FASR	Audio Pres.	Video Pres.	FOSR
RAVDESS	FusionBreak	93.81 ± 4.44	94.77 ± 3.87	86.72 ± 7.85	77.09 ± 7.92
	FusionBreak (detector-aware)	92.98 ± 4.95	94.55 ± 4.27	86.51 ± 7.99	75.99 ± 8.80
	PGD	100.00 ± 0.00	1.30 ± 2.88	2.84 ± 3.23	0.00 ± 0.00
	CE-only selection	100.00 ± 0.00	67.23 ± 10.78	19.85 ± 8.75	10.02 ± 6.99
	Audio-only FusionBreak	72.05 ± 13.54	93.59 ± 3.81	100.00 ± 0.00	65.64 ± 11.54
	Video-only FusionBreak	29.49 ± 8.44	100.00 ± 0.00	79.54 ± 6.57	9.02 ± 7.06
CREMA-D	FusionBreak	100.00 ± 0.00	99.94 ± 0.13	97.73 ± 1.73	97.67 ± 1.75
	FusionBreak (detector-aware)	100.00 ± 0.00	99.81 ± 0.33	97.83 ± 1.84	97.64 ± 1.91
	PGD	100.00 ± 0.00	5.52 ± 12.10	0.00 ± 0.00	0.00 ± 0.00
	CE-only selection	100.00 ± 0.00	89.42 ± 3.30	22.70 ± 5.27	18.68 ± 5.05
	Audio-only FusionBreak	87.41 ± 17.62	98.06 ± 1.73	100.00 ± 0.00	85.47 ± 18.75
	Video-only FusionBreak	100.00 ± 0.00	100.00 ± 0.00	98.25 ± 1.52	98.25 ± 1.52

Detector-aware optimization reduces the mean detection rate from 11.21% to 8.39% on RAVDESS and from 26.28% to 19.23% on CREMA-D. In comparison, PGD has mean detection rates of 32.86% and 46.49%, respectively. The largest adaptive reduction occurs for Mahalanobis detection, from 21.25% to 12.63% on RAVDESS and from 55.27% to 37.45% on CREMA-D. Detection behavior differs across detectors, and LID is not included in the adaptive objective. Full per-detector results are reported in Appendix D.

We also evaluate a structural defense that flags a sample when the audio and video predictions agree, but the fused prediction differs. By definition, this rule detects every successful fusion-only example. However, such disagreement can also occur on clean inputs, so the defense can produce false positives. Its clean-test false-positive rate is $1.77 \pm 1.46\%$ on RAVDESS and $4.95 \pm 1.48\%$ on CREMA-D. Therefore, although this consistency check is effective against FusionBreak, it introduces a considerable false-positive trade-off on benign samples.

6 Ablation Studies

Component ablations show that explicit unimodal preservation and decision-aware candidate selection are the most important components. Removing the preservation objective reduces FOSR from 77.09% to 10.82% on RAVDESS and from 97.67% to 33.07% on CREMA-D, while retaining only the final optimization iterate reduces FOSR to 42.97% and 24.25%, respectively. Removing the fusion-margin and output-consistency terms has substantially smaller effects. Full component results are reported in Appendix E.

Table 3 examines sensitivity to the attack configuration. Increasing the number of restarts and optimization steps improves FOSR on both datasets, but the gains diminish. For example, increasing the number of steps from 40 to 80 improves FOSR by 6.60 percentage points on RAVDESS and 2.42 percentage points on CREMA-D, whereas increasing from 80 to 120 adds only 2.81 and 0.71 percentage points, respectively.

The effect of perturbation budget depends on the dataset. On RAVDESS, larger budgets increase both FASR and FOSR. On CREMA-D, FASR is already 100% at half the default budget, while FOSR decreases as the budget increases. This suggests that once fused attack success is saturated, additional perturbation can make preservation of the unimodal decisions more difficult.

7 Discussion and Future Work

FusionBreak shows that conventional fused attack success and fusion-only failure capture different behaviors. Standard PGD can fool the fused model while also disrupting the unimodal predictions, whereas FusionBreak explicitly searches for cases in which the unimodal decisions remain correct.

Table 3: Sensitivity to perturbation budget, number of restarts, and optimization steps. Values are mean $\pm$ standard deviation across three seeds and four fusion schemes. Default settings use $\epsilon_a = 0.08$, $\epsilon_v = 0.06$, $R = 10$, and $T = 80$. All values are percentages.

		RAVDESS		CREMA-D	
Setting	Value	FASR	FOSR	FASR	FOSR
Budget	$0.5\times$	68.17 ± 13.24	60.06 ± 14.00	100.00 ± 0.00	99.92 ± 0.19
	Default	93.81 ± 4.44	77.09 ± 7.92	100.00 ± 0.00	97.67 ± 1.75
	$1.5\times$	99.03 ± 1.46	86.86 ± 7.08	100.00 ± 0.00	95.51 ± 3.23
Restarts	1	87.72 ± 5.11	69.26 ± 7.91	100.00 ± 0.00	95.73 ± 0.77
	5	93.58 ± 4.40	74.73 ± 8.74	100.00 ± 0.00	97.10 ± 1.82
	10	93.81 ± 4.44	77.09 ± 7.92	100.00 ± 0.00	97.67 ± 1.75
Steps	40	90.90 ± 5.65	70.49 ± 9.01	100.00 ± 0.00	95.25 ± 1.99
	80	93.81 ± 4.44	77.09 ± 7.92	100.00 ± 0.00	97.67 ± 1.75
	120	94.78 ± 3.76	79.90 ± 7.45	100.00 ± 0.00	98.38 ± 1.35

This distinction suggests that multimodal robustness evaluation should consider the behavior of both the component classifiers and the final fused prediction. The single-modality controls further show that fusion-only vulnerability does not follow one universally dominant modality. Audio-only and video-only attacks behave differently across RAVDESS and CREMA-D, suggesting that the relevant attack path depends on how modality-specific information contributes to the fused decision.

Detector-aware optimization reduces detection by the generic detectors considered here while largely preserving fusion-only effectiveness. However, the hard-label consistency rule directly identifies successful fusion-only examples whenever the unimodal predictions agree on the correct class but the fused prediction differs. This suggests that fusion-specific consistency checks may complement generic adversarial detectors, although such checks may be less reliable when unimodal predictions naturally disagree.

The term *fusion-only failure* refers to the observed decision condition, not to a claim that the fusion module is the only component affected by the perturbation. Although the unimodal predicted labels remain correct, their internal representations and confidence values may change and can influence the fused prediction. Therefore, FusionBreak isolates failures that occur despite preserved unimodal decisions rather than proving that the fusion operator alone is causally responsible.

FOSR should be interpreted together with the eligible rate. Because $\mathcal{E}$ contains only samples for which the clean audio, visual, and fused predictions are all correct, FOSR measures the vulnerability of initially correct multimodal decisions rather than attack success over the full test distribution. Therefore, a high FOSR on a model with a small eligible set has a different interpretation from the same FOSR on a model with broad clean eligibility. Table 1 reports these eligible rates for context.

Our evaluation is limited to white-box attacks on two audio-visual emotion-recognition datasets and four fusion schemes. Future work can extend FusionBreak to other modality combinations and tasks, transformer-based or larger multimodal models, and transfer-based or black-box threat models. Another direction is to develop fusion-aware defenses that reason jointly about unimodal representations, predictions, and the final fused decision.

8 Conclusion

We introduced FusionBreak and FOSR to study adversarial failures in which the fused classifier becomes incorrect while the unimodal classifiers remain correct. Experiments across two datasets and four fusion schemes show that high FASR does not necessarily imply a fusion-only failure, highlighting the need to evaluate unimodal preservation explicitly. Detector-aware optimization reduces detection by several generic adversarial detectors while largely preserving fusion-only attack effectiveness, motivating further study of fusion-aware defenses. Overall, these results show that multimodal robustness evaluations should jointly consider fused attack success, unimodal preservation, and fusion-specific consistency.

AI USE STATEMENT

We used generative AI tools to assist with drafting and polishing parts of the manuscript, literature retrieval and verification, code implementation and debugging, and the creation or refinement of schematic visualizations. We independently reviewed and verified all AI-assisted work and take responsibility for the final text, claims, code, figures, and experimental results.

ETHICS STATEMENT

This work studies adversarial vulnerabilities in multimodal learning systems. Attackers could potentially use the proposed attack to identify or exploit weaknesses in deployed multimodal models. Our experiments are intended to support robustness evaluation by showing that a fused model can fail even when its unimodal predictions remain correct. We evaluate the attack only in controlled, offline experimental settings and do not target deployed systems or real users.

The experiments use the publicly available RAVDESS and CREMA-D datasets and do not involve new data collection or interaction with human participants. We also evaluate adversarial detection and a fusion-specific consistency check to characterize possible defenses against the studied failure mode.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide the model, training, attack, detector, and evaluation details needed to reproduce the experiments. We have also provided anonymized source code, together with instructions for running the project, in the Supplementary Material section of the submission portal. The datasets used in this work are publicly available and should be downloaded from their respective original sources.

REFERENCES

Aiswarya Akumalla, Seth Haney, and Maksim Bazhenov. Contextual fusion for adversarial robustness. *arXiv preprint arXiv:2011.09526*, 2020.

Eugene Bagdasaryan, Rishi Jha, Vitaly Shmatikov, and Tingwei Zhang. Adversarial illusions in {Multi-Modal} embeddings. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 3009–3025, 2024.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Ivan Evtimov, Russel Howes, Brian Dolhansky, Hamed Firooz, and Cristian Canton Ferrer. Adversarial evaluation of multimodal models under realistic gray box assumption. *arXiv preprint arXiv:2011.12902*, 2020.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26866–26875, 2024. doi: 10.1109/CVPR52733.2024.02538.

Jiwei Guan, Tianyu Ding, Longbing Cao, Lei Pan, Chen Wang, and Xi Zheng. Probing the robustness of vision-language pretrained models: A multimodal adversarial attack approach. *arXiv preprint arXiv:2408.13461*, 2024.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Taewan Kim and Joydeep Ghosh. On single source robustness in deep fusion models. *Advances in Neural Information Processing Systems*, 32, 2019.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

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

Moru Liu, Hao Dong, Olga Fink, and Mario Trapp. Adaptive confidence regularization for multimodal failure detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15850–15859, 2026.

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*, 13(5):e0196391, 2018. doi: 10.1371/journal.pone.0196391.

Haiyan Long, Hai Chen, Mengyao Xu, Chonghao Zhang, and Fulan Qian. Crafting transferable adversarial examples against 3d object detection. *IET Computer Vision*, 19(1):e70011, 2025. doi: https://doi.org/10.1049/cvi2.70011.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 102–111, 2023.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Michael E. Houle, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Brandon Mckinzie, Vaishaal Shankar, Joseph Yitan Cheng, Yinfei Yang, Jonathon Shlens, and Alexander T Toshev. Robustness in multimodal learning under train-test modality mismatch. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24291–24303. PMLR, 2023.

Mohammad Meymani, Roozbeh Razavi-Far, Arash Vashagh, and Battista Biggio. Defense against adversarial attacks: Foundations, strategies, and future directions. *Preprints*, 2026. doi: 10.20944/preprints202607.1022.v1.

Wonseok Park, Qi Alfred Chen, and Zhanpeng Mao. Crafting adversarial examples on 3d object detection sensor fusion models. In *Proceedings of the CVPR 2020 Workshop on Adversarial Machine Learning in Computer Vision*, 2020.

Maya Srikanth, Run Chen, and Julia Hirschberg. Mixed signals: Understanding model disagreement in multimodal empathy detection. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1978–1991, 2025.

Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5601–5611, 2021.

James Tu, Huichen Li, Xinchen Yan, Mengye Ren, Yun Chen, Ming Liang, Eilyan Bitar, Ersin Yumer, and Raquel Urtasun. Exploring adversarial robustness of multi-sensor perception systems in self driving. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1013–1024. PMLR, 2022.

Arash Vashagh, Roozbeh Razavi-Far, Mohammad Meymani, and Battista Biggio. Recent advances in adversarial attacks on model utility, privacy, and explainability: A comprehensive survey. *TechRxiv*, 2026(0305), 2026. doi: 10.36227/techrxiv.177272853.30003431/v1.

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

Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1722–1740, 2024a. doi: 10.1109/SP54263.2024.00102.

Yanting Wang, Hongye Fu, Wei Zou, and Jinyuan Jia. Mmcert: Provable defense against adversarial attacks to multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24655–24664, 2024b.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Proceedings 2018 Network and Distributed System Security Symposium*, 2018. doi: 10.14722/ndss.2018.23198.

Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3340–3349, 2021.

Jehyeok Yeon, Hyeonjeong Ha, Qiusi Zhan, and Heng Ji. Securing multimodal AI through internal information decomposition. In *Forty-third International Conference on Machine Learning*, 2026.

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

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*, MM '22, pp. 5005–5013. Association for Computing Machinery, 2022. doi: 10.1145/3503161.3547801.

Junrui Zhang, Xinyu Zhao, Jie Peng, Chenjie Wang, Jianmin Ji, and Tianlong Chen. Vulnerability-aware robust multimodal adversarial training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(33):28265–28273, 2026a. doi: 10.1609/aaai.v40i33.40054.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Qinghua Hu, Cheng Deng, Cai Xu, Jie Wen, Di Hu, and Changqing Zhang. Multimodal fusion on low-quality data: A comprehensive survey. *Information Fusion*, 135:104437, 2026b. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2026.104437.

Zeliang Zhang, Susan Liang, Daiki Shimada, and Chenliang Xu. Rethinking audio-visual adversarial vulnerability from temporal and modality perspectives. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Additional Experimental Details

This appendix provides additional implementation details for the experiments reported in the main paper, including preprocessing, training, attack optimization, and detector settings.

A.1 Preprocessing and Training

For each seed, RAVDESS uses random 70/15/15% train/validation/test splits, while CREMA-D uses class-stratified 70/10/20% splits.

For RAVDESS, audio is resampled to 16 kHz and fixed to three seconds by center cropping or zero padding. We compute a log-magnitude STFT using $n_{\text{fft}} = 512$, a 512-sample Hann window, and hop length 256, resize it to 64×64, and standardize each representation. For video, 12 frames are uniformly sampled from approximately the central 80% of each clip, converted to grayscale after a slightly upper-centered square crop, resized to 64×64, and normalized to $[-1, 1]$.

For CREMA-D, audio is resampled to 16 kHz using polyphase resampling, amplitude-normalized, and zero padded or truncated to three seconds. We compute $\log(1 + |\text{STFT}|)$ using $n_{\text{fft}} = 512$, a 400-sample Hann window, and hop length 160, and resize the result to 64×128. Video uses 12 uniformly sampled grayscale frames resized to 64×64 with values in $[0, 1]$.

The audio encoder uses convolutional layers with 32, 64, and 128 channels, and both unimodal branches produce 128-dimensional representations. Models are trained with cross-entropy and AdamW, batch size 64, and weight decay 10^{-4}. RAVDESS uses learning rates 7×10^{-4} for unimodal training and 3×10^{-4} for fusion training; CREMA-D uses 10^{-3} for both. Unimodal and fusion models are trained for at most 40 and 35 epochs, respectively, with early-stopping patience 8 and validation-accuracy checkpoint selection.

RAVDESS training additionally uses Gaussian audio noise ($\sigma = 0.015$), time/frequency masking of up to 12% of each dimension, horizontal video flipping with probability 0.5, $\pm 10\%$ brightness jitter, and video noise with $\sigma = 0.015$. The unimodal backbones are frozen during RAVDESS fusion training. For CREMA-D, the copied unimodal backbones are fine-tuned jointly with the fusion model and no additional training augmentation is applied. Mixed-precision training is enabled when a GPU is available.

A.2 Attack and Detector Settings

Table 4 summarizes the threat model, optimization procedure, perturbation constraints, objective weights, candidate-selection strategy, and detector settings used in the main experiments.

The first restart begins from the clean input, while subsequent restarts are randomly initialized within the allowed perturbation sets. Candidates are evaluated after every optimization step and ranked according to the fusion-only objective, prioritizing fused misclassification with both unimodal decisions preserved. Fused cross-entropy and unimodal margins are used as tie-breakers.

For detector-aware optimization, MSP, Mahalanobis distance, and feature squeezing are incorporated into the adaptive objective. Feature squeezing uses a straight-through approximation through its quantization step during optimization, while the original detector is used for evaluation. LID is evaluated separately but is not included in the adaptive objective because its nearest-neighbor-based score is not directly suited to the gradient-based optimization used here. Detector thresholds are calibrated on clean validation data.

We apply perturbations to the processed audio and video inputs. Projection enforces the corresponding L_∞ budget. CREMA-D video is additionally clipped to its $[0, 1]$ input range.

The generic detectors are fitted using up to 2,500 clean training samples. MSP uses the anomaly score $1 - \max_c p_c$. Mahalanobis distance and LID operate on the L_2-normalized pre-classifier fused representation. Mahalanobis uses class-conditional means and a shared covariance matrix regularized by $10^{-3}I$, with its pseudoinverse used as the precision matrix. LID uses $k = 20$ nearest neighbors. Feature squeezing applies 5-bit quantization followed by 3×3 average pooling to both modalities and scores the L_1 difference between the original and squeezed fused softmax outputs.

Detector thresholds are set to the 95th percentile of clean-validation scores, corresponding to a target 5% false-positive rate. For the detector-aware objective, q_j is the standard deviation of detector j's clean-validation scores; values below 10^{-6} are replaced by $\max(|\tau_j|, 1)$. MSP, Mahalanobis, and feature squeezing use weights 0.5, 3.0, and 5.0, respectively, with $\lambda_D = 2.0$. Feature squeezing uses a straight-through approximation to its quantization operation during optimization. LID is evaluated but is not included in the adaptive objective.

Table 4: Attack and detector settings used in the main experiments.

Setting	Parameter	Value
Threat model	Access	White-box
	Attack type	Untargeted
	Perturbed modalities	Audio and video
	Perturbation norm	L_∞
Optimization	Steps / Restarts	80 / 10
	Initialization	Clean first; random thereafter
	Candidate evaluation	Every optimization step
Perturbation	ϵ_a/ϵ_v	0.08 / 0.06
	α_a/α_v	0.004 / 0.003
Objective	λ_F/λ_M	0.5 / 5.0
	λ_K/ρ	0.25 / 0.05
	λ_D	2.0
Candidate selection	Primary criterion	Fusion-only success
	Secondary criterion	Preserve one unimodal decision
Detector-aware attack	Optimized detectors	MSP / Mahalanobis / FS
	Detector weights	0.5 / 3.0 / 5.0
	Evaluation-only detector	LID
Calibration	Threshold data	Clean validation set
	Target FPR	5%

B Per-Scheme Attack Results

Table 5 separates FusionBreak performance by fusion scheme. All four designs exhibit fusion-only failures on both datasets. The RAVDESS interaction model has the lowest mean FOSR at 74.38%, while concatenation has the highest at 80.83%. The results are closer on CREMA-D, where mean FOSR lies between 97.13% and 98.44%. RAVDESS Gated-II has the largest variation across seeds, but its mean FOSR remains 75.10%. These results support the finding that the failure mode is present across the tested fusion designs.

Table 5: FusionBreak results by fusion scheme. FASR and FOSR are mean $\pm$ standard deviation in percent across seeds 42, 43, and 44. Eligible n is the total number of attacked clean samples across the three seeds.

Dataset	Fusion scheme	Eligible n	FASR	FOSR
RAVDESS	Concatenation	100	94.99 ± 1.72	80.83 ± 6.96
	Interaction	99	89.25 ± 4.71	74.38 ± 4.48
	Gated-I	100	95.28 ± 4.11	78.06 ± 5.02
	Gated-II	100	95.70 ± 4.93	75.10 ± 14.57
CREMA-D	Concatenation	948	100.00 ± 0.00	97.13 ± 2.88
	Interaction	890	100.00 ± 0.00	97.43 ± 1.07
	Gated-I	863	100.00 ± 0.00	97.70 ± 2.00
	Gated-II	897	100.00 ± 0.00	98.44 ± 1.43

C Computational Cost

FusionBreak performs $R \times T$ optimization steps per sample, where R is the number of random restarts and T is the number of steps per restart. Under the default configuration of 10 restarts and 80 steps, the attack performs 800 gradient-based optimization steps per eligible sample. Candidates are evaluated after every step, and the attack retains the highest-priority intermediate candidate according to the selection procedure described in Section 3.

The detector-aware variant uses the same number of optimization steps but incurs additional per-step computation because MSP, Mahalanobis distance, and feature squeezing are evaluated during optimization. LID does not contribute to the adaptive optimization cost because it is used only for evaluation. The sensitivity results in Table 3 show diminishing returns beyond 80 steps and 10 restarts, motivating the default configuration as a trade-off between attack effectiveness and computational cost.

D Per-Detector Detection Results

Table 6 reports the complete detector-specific results corresponding to the summary reported in the main paper.

Table 6: Detection rates across the 12 configurations. Detection rate is the fraction of attacked eligible samples whose score exceeds the threshold calibrated on clean validation data. All values are percentages.

Dataset	Detector	FusionBreak	Adaptive	PGD
RAVDESS	Feature squeezing	6.08	5.87	62.89
	LID	9.72	7.91	3.99
	Mahalanobis	21.25	12.63	64.57
	MSP	7.81	7.14	0.00
	Mean	**11.21**	**8.39**	**32.86**
CREMA-D	Feature squeezing	28.47	25.75	50.30
	LID	19.96	12.42	49.97
	Mahalanobis	55.27	37.45	85.67
	MSP	1.41	1.29	0.00
	Mean	**26.28**	**19.23**	**46.49**

Detector-aware optimization reduces the mean detection rate from 11.21% to 8.39% on RAVDESS and from 26.28% to 19.23% on CREMA-D. The largest reduction occurs for Mahalanobis detection, while feature squeezing and MSP change only modestly. PGD is more detectable on average, although individual detectors behave differently. In particular, MSP assigns a 0% detection rate to PGD on both datasets.

E Additional Component Ablations

We report the full component ablation results in Table 7. All variants use the same final checkpoints, data splits, and eligible samples within each configuration. Only the component or selection rule named by the variant is changed. Results are averaged across three seeds and four fusion schemes.

Table 7: Full component ablations across three seeds and four fusion schemes. Each entry is mean ± standard deviation in percent. The CE-only row is from the ablation sweep and is separate from the equal-search CE-only control reported in Table 2.

Dataset	Variant	FASR	Audio Pres.	Video Pres.	FOSR
RAVDESS	Fusion CE only	100.00±0.00	67.23±10.78	19.85±8.75	10.02±6.99
	Fusion CE + margin	100.00±0.00	66.48±10.76	22.78±11.00	10.83±7.97
	Without preservation	100.00±0.00	66.97±11.05	23.05±10.78	10.82±7.33
	Without fusion margin	92.17±6.39	94.33±5.48	87.33±9.56	75.72±7.90
	Without output consistency	94.89±4.41	94.82±4.32	83.02±9.98	74.94±9.56
	Final-iterate selection	77.11±5.78	79.23±8.33	77.19±12.76	42.97±12.79
	Full FusionBreak	93.81±4.44	94.77±3.87	86.72±7.85	77.09±7.92
CREMA-D	Fusion CE only	100.00±0.00	89.50±3.71	22.52±5.39	18.70±5.28
	Fusion CE + margin	100.00±0.00	89.85±3.66	22.84±5.99	18.79±5.55
	Without preservation	100.00±0.00	55.51±11.63	77.08±14.95	33.07±10.26
	Without fusion margin	99.90±0.20	99.97±0.09	97.97±1.35	97.85±1.45
	Without output consistency	100.00±0.00	99.92±0.20	97.42±1.79	97.34±1.83
	Final-iterate selection	100.00±0.00	48.11±9.11	52.25±17.08	24.25±10.50
	Full FusionBreak	100.00±0.00	99.94±0.13	97.73±1.73	97.67±1.75

The full ablation results confirm that explicit unimodal preservation is the main source of fusion-only success. Removing the preservation objective reduces FOSR from 77.09% to 10.82% on RAVDESS and from 97.67% to 33.07% on CREMA-D. Fusion CE alone and Fusion CE with the fused-margin term also produce low FOSR despite reaching 100% FASR.

Decision-aware candidate selection is also important. Keeping only the final optimization iterate reduces FOSR from 77.09% to 42.97% on RAVDESS and from 97.67% to 24.25% on CREMA-D. This indicates that successful intermediate states can be lost when optimization continues after a fusion-only failure has already been found.

The fusion-margin and output-consistency terms have smaller effects. Removing the fusion margin changes FOSR from 77.09% to 75.72% on RAVDESS and from 97.67% to 97.85% on CREMA-D. Removing output consistency changes FOSR to 74.94% and 97.34%, respectively. These terms provide more modest support than the preservation objective and decision-aware selection.