SoundBreak: A Systematic Study of Audio-Only Adversarial Attacks on Trimodal Models

Aafiya Hussain[♡], **Gaurav Srivastava**[♡], **Alvi Md Ishmam**[♡], **Zaber Hakim**[♡], **Chris Thomas**[♡]
[♡]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA,
[♡](aafiyahussain, gks, alvi, zaberhakim666, christhomas)@vt.edu

Leaderboard: aafiya-h.github.io/soundbreak

Abstract

Multimodal foundation models that integrate audio, vision, and language achieve strong performance on reasoning and generation tasks, yet their robustness to adversarial manipulation remains poorly understood. We study a realistic and underexplored threat model: **untargeted, audio-only adversarial attacks** on trimodal audio–video–language models. We analyze six complementary attack objectives that target different stages of multimodal processing, including audio encoder representations, cross-modal attention, hidden states, and output likelihoods. Across four state-of-the-art models and multiple benchmarks, we show that audio-only perturbations can induce severe multimodal failures, achieving up to **96% attack success rate.** We further show that attacks can be successful at low perceptual distortions (LPIPS ≤ 0.08, SI-SNR ≥ 0 dB) and benefit more from extended optimization than increased data scale. We evaluate the feasibility of these attacks under physically realistic conditions by incorporating room impulse response (RIR) modeling, showing that audio-only perturbations remain effective under environmental transformations and thus highlight the practical risk of single-modality attacks in real-world multimodal systems. Transferability across models and encoders remains limited, while speech recognition systems such as Whisper primarily respond to perturbation magnitude, achieving **>97% attack success** under severe distortion. These results expose a previously overlooked single-modality attack surface in multimodal systems and motivate defenses that enforce cross-modal consistency.

1 Introduction

Recent progress in multimodal large language models has intensified research on their susceptibility to adversarial attacks across multiple modalities (Liu et al., 2025b). While these models demonstrate strong capabilities in integrating audio, vision, and language, prior work has revealed critical vulnerabilities in CLIP-based and vision-language systems (Zhang et al., 2025c; Mei et al., 2025; Xu et al., 2024), with growing interest in audio-domain attacks (Raina et al., 2024). Despite these advances, robustness against adversarial manipulation remains largely unexplored in settings where the attacker controls only a single modality within a multimodal input pipeline.

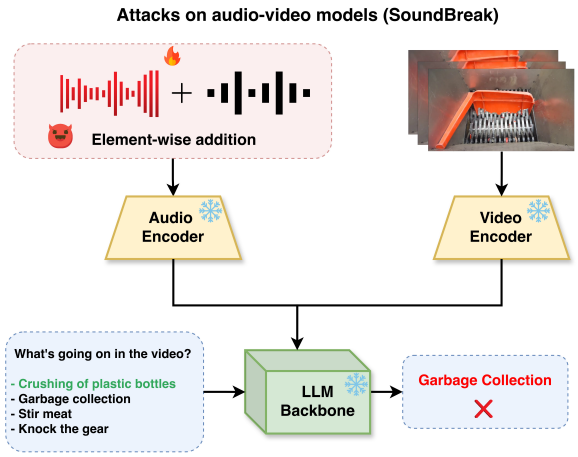

Figure 1: Audio-only adversarial attacks on audio–video–language models. An additive perturbation applied solely to the audio stream propagates through the model, resulting in incorrect outputs.

Adversarial attacks on multimodal models broadly fall into four categories: *prompt injection*, *jailbreaks*, *data poisoning*, and *adversarial perturbations* (Liu et al., 2025b; Jiang et al., 2025; Shayegani et al., 2023; Lu et al., 2025). Jailbreak attacks bypass safety mechanisms through prompt engineering (Yi et al., 2024; Liu et al., 2024) or gradient-based optimization (Wang et al., 2024), while prompt injection embeds malicious instructions directly into inputs such as images or audio (Bagdasaryan et al., 2023). Data poisoning and backdoor attacks compromise training pipelines (Walmer et al., 2022). Among these, ad-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27635–27663

versarial perturbations pose a particularly unique cross-modal threat: by modifying a single input modality, they can indirectly influence model reasoning across all modalities without explicitly altering them (Kang et al., 2025a). This attack surface remains insufficiently understood.

Existing work has predominantly focused on unimodal or bimodal settings. Vision-language attacks demonstrate that carefully crafted image perturbations can manipulate model outputs across downstream tasks (Zhang et al., 2025c; Mei et al., 2025; Hao et al., 2025; Zhang et al., 2025a, 2022). Separately, audio-domain attacks show that speech recognition and translation systems can be silenced, redirected, or manipulated through acoustic adversarial examples (Raina et al., 2024; Ma et al., 2025; Sadasivan et al., 2026; Liu et al., 2023). However, these approaches treat modalities largely in isolation, failing to capture the complex cross-modal dependencies inherent to modern multimodal foundation models.

A small number of multimodal attack studies rely on simultaneous manipulation of multiple input channels (Mustakim et al., 2025; Wang et al., 2025; Tian and Xu, 2021) or require access to training data for backdoor insertion (Walmer et al., 2022; Han et al., 2024; Liang et al., 2024; Yu et al., 2025; Liang et al., 2025). These assumptions limit practical relevance: coordinating attacks across multiple modalities is technically challenging and easily detectable, while training-time attacks assume access to training data. Moreover, targeted attacks that aim to induce specific malicious outputs often introduce conspicuous behavioral anomalies that can be flagged by monitoring systems.

In this work, we investigate a more realistic and underexplored threat model: **untargeted, audio-only adversarial attacks on audio-video-language models**. We emphasize that our attacks are not per-sample optimizations, but shared perturbations learned through optimization over training data. We demonstrate that an attacker who controls only the audio channel can systematically degrade multimodal reasoning through gradient-based optimization, without modifying visual or textual inputs. This threat model is particularly concerning because audio manipulation is easier to deploy in real-world settings via compromised microphones, environmental speakers, or transmission channels, while being significantly harder to detect than visual perturbations (Choi et al., 2024). We further extend this analysis to physically realistic settings by incorporating room impulse response (RIR) modeling, allowing us to simulate environmental effects such as reverberation and signal propagation. Our results show that audio-only perturbations remain effective under such transformations, reinforcing the practical relevance of this threat model.

Our contributions are fourfold. ***First,*** we propose six complementary audio-based adversarial objectives that target different stages of multimodal processing, including encoder representations, attention allocation, hidden states, and output likelihoods. ***Second,*** we conduct extensive evaluations across four state-of-the-art audio-visual models on standardized multimodal benchmarks, revealing model-specific vulnerabilities and limited cross-model transferability. ***Third,*** we analyze how attack effectiveness depends on training duration, data efficiency, perturbation magnitude, and perceptual distortion, yielding practical insights into both attack construction and defensive design. ***Fourth,*** We evaluate the robustness of audio-only adversarial attacks under physically realistic conditions using room impulse response (RIR) modeling, demonstrating their effectiveness beyond digital settings.

2 Related Work

Foundations of Adversarial Attacks. The discovery of adversarial examples fundamentally challenged prevailing assumptions about neural network robustness (Szegedy et al., 2013; Nguyen et al., 2015). Early investigations highlighted the structural fragility of deep models under small, adversarially chosen perturbations (Papernot et al., 2016), while the Fast Gradient Sign Method demonstrated how approximately linear behavior in high-dimensional spaces gives rise to these vulnerabilities (Goodfellow et al., 2015). Subsequent work introduced increasingly precise optimization-based attacks, ranging from DeepFool's geometric formulation (Moosavi-Dezfooli et al., 2016) to the Carlini-Wagner attack (Carlini and Wagner, 2017), establishing that adversarial robustness cannot be achieved through superficial defenses alone. More recently, these principles have been successfully extended to large language models via projected gradient descent on relaxed input representations, enabling efficient and scalable attacks in previously intractable settings (Geisler et al., 2024).

Vision-Centric Multimodal Attacks. With the rise of vision-language models, adversarial research has increasingly focused on vulnerabilities at the vision-text interface. Query-agnostic attacks demonstrate that a single adversarial image can generalize across multiple downstream questions by disrupting vision-language alignment (Zhang et al., 2025c). Related work shows that targeting only the vision encoder suffices to degrade performance across diverse tasks, highlighting encoder representations as a critical attack surface (Mei et al., 2025). Scenario-aware and multi-loss attacks further amplify these effects, achieving high jailbreak success rates on commercial systems (Hao et al., 2025). Scalability has been addressed through self-supervised pretraining strategies that enable transferable any-to-any attacks across tasks and prompts (Zhang et al., 2025b). However, these approaches primarily rely on visual perturbations and do not examine how vulnerabilities propagate across more than two modalities.

Audio-Domain Adversarial Techniques. The audio modality introduces distinct adversarial challenges due to temporal structure, perceptual constraints, and physical-world variability. Universal acoustic segments can suppress or terminate speech recognition by mimicking control tokens (Raina et al., 2024), while targeted perturbations enable conditional manipulation based on speaker identity or content (Ma et al., 2025). Over-the-air attacks demonstrate that adversarial audio can remain effective despite environmental noise and reverberation (Sadasivan et al., 2026). Recent studies extend these ideas to speech translation systems, showing that both imperceptible perturbations and adversarial music can induce malicious translations (Liu et al., 2025a). Defense mechanisms such as SPIRIT propose activation patching to mitigate such attacks, but also expose the trade-off between robustness and model utility (Djanibekov et al., 2025). AdvWave further demonstrates that gradient shattering in audio models can be overcome through dual-phase optimization while preserving perceptual naturalness (Kang et al., 2025b). Notably, these works focus primarily on speech-centric models and do not address multimodal reasoning.

Audio-Only Attacks in Multimodal Systems. Table 1 summarizes how existing adversarial attacks differ in their threat models and assumptions. Prior work largely focuses on vision-language settings, unimodal speech models, or attacks that require coordinated manipulation of multiple modalities. As a result, these approaches either assume access to multiple input channels, target specific outputs, or operate outside truly multimodal reasoning settings. In contrast, SOUNDBREAK is the first to study *untargeted, audio-only attacks* in trimodal audio-video-language models under a realistic threat model, while providing systematic analysis of encoder-space, attention-based, and hidden-state level vulnerabilities. This positioning highlights a previously unexplored attack surface in multimodal systems.

3 SOUNDBREAK Setup

3.1 Problem Formulation

Let $M_\theta : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{Q} \to \mathcal{P}(\mathcal{A})$ denote an audio-video model that consumes a natural-language question $q \in \mathcal{Q}$ together with synchronized audio $x_a \in \mathcal{X}_a$ (of length T) and video $x_v \in \mathcal{X}_v$, and returns a distribution $p(\cdot \mid x_a, x_v, q; \theta)$ over answers, where θ represents the model parameters. We study a white-box, audio-only additive adversary that seeks to alter the model's answer in an *untargeted* fashion. The adversary injects a perturbation $\delta \in \mathbb{R}^T$ to produce $\tilde{x}_a = x_a + \delta$ subject to a feasibility set

$$\mathcal{C} = \{\delta : \|\delta\|_\infty \le \varepsilon\}, \tag{1}$$

which captures the ℓ_∞ norm constraint with ε denoting the attack budget.

Under the white-box assumption, the attacker has full access to θ, model outputs, and internal differentiable representations. The attacker may therefore evaluate gradients $\nabla_\delta \mathcal{L}^{(i)}$ for each chosen loss index $i \in \{1, \dots, k\}$. We define the iterative refinement of the additive perturbation by

$$\delta^{(k+1)} = \Pi_{\mathcal{C}}\Big(\delta^{(k)} - \eta\, \nabla_\delta\, \mathcal{L}^{(i)}\big(p(\cdot \mid x_a + \delta^{(k)}, x_v, q; \theta), p(\cdot \mid x_a, x_v, q; \theta)\big)\Big), \tag{2}$$

where $\delta^{(0)} \in \mathcal{C}$, $\Pi_{\mathcal{C}}$ denotes projection onto the feasible set $\mathcal{C}$, and $\eta > 0$ is the step size. After K iterations, the attack uses $\tilde{x}_a = x_a + \delta^{(K)}$ as the perturbed audio input. We further evaluate this attack on other models in a black-box setting using $\tilde{x}_a$ as the new audio input.

3.2 Attack Methodology

We design our attack methodology to systematically stress-test trimodal audio-video-language

	QAVA	VEAttack	Muting Whisper	AdvWave	Multimodal Attacks	SoundBreak
Targets Audio Modality	✗	✗	✓	✓	✓	✓
Single-Modality Control	✗	✓	✓	✓	✗	✓
Trimodal Setting (A+V+L)	✗	✗	✗	✗	✓	✓
Untargeted Attack Objective	✓	✓	✓	✗	✗	✓
Query-Agnostic / Task-Agnostic	✓	✓	✓	✗	✗	✓
Realistic Threat Model	△	△	✓	△	✗	✓
Attack Interpretability Analysis	✗	✗	✗	△	✗	✓

Table 1: Comparison of prior adversarial attack methods on multimodal models. Existing work largely focuses on vision-language or unimodal settings, targeted objectives, or simultaneous multi-modal manipulation. SoundBreak uniquely studies untargeted, audio-only attacks in trimodal models with systematic analysis of encoder-space, attention-based, and hidden-state vulnerabilities, as well as transferability across models and encoders. △ denotes partial support. Multimodal Attacks refer to the attack from (Mustakim et al., 2025).

models under deviations in the audio channel while keeping visual and textual inputs fixed. We employ a diverse set of adversarial losses that probe complementary stages of the audio-to-output pipeline, including representation learning, cross-modal interaction, and internal attention dynamics. This multi-view formulation allows us to disentangle which computational components are most susceptible to audio-only perturbations and how failures propagate across modalities. Lastly, all attacks optimize a shared audio perturbation rather than per-sample adversarial noise, enabling analysis of systematic model vulnerabilities. Detailed mathematical formulations and algorithms are in appendices C, D.

3.2.1 Negative Language Modeling Loss

A direct way to induce adversarial failure is to reduce the model's confidence in its original answer under perturbed audio input. We implement this by defining an objective derived from the language modeling loss, which serves as a scalar proxy for answer likelihood in multimodal generation.

Let $a^\star \in \mathcal{A}$ denote the ground-truth answer associated with input (x_a, x_v, q). The standard language modeling loss under perturbed audio $\tilde{x}_a = x_a + \delta$ is given by

$$\mathcal{L}_{\text{LM}}\big(p(\cdot \mid \tilde{x}_a, x_v, q; \theta), a^\star\big) = -\log p(a^\star \mid \tilde{x}_a, x_v, q; \theta). \tag{3}$$

To explicitly suppress the probability assigned to the correct answer, we define the adversarial objective as

$$\mathcal{L}_{\text{negLM}} = -\mathcal{L}_{\text{LM}}. \tag{4}$$

This formulation is motivated by recent work showing that directly optimizing inputs against the language modeling objective is an effective mechanism for attacking large language models using projected gradient descent (Geisler et al., 2025).

3.2.2 Encoder-Based Cosine Similarity Loss

To directly probe vulnerabilities in audio representation learning, we target the audio encoder rather than downstream language components. Let $f_a : \mathcal{X}_a \to \mathbb{R}^d$ denote the audio encoder of M_θ, which maps an input waveform to a d-dimensional embedding. We define an encoder-space adversarial objective that minimizes the cosine similarity between the embeddings of clean and perturbed audio:

$$\mathcal{L}^{(\cos)}(\tilde{x}_a, x_a) = \frac{f_a(\tilde{x}_a) \cdot f_a(x_a)}{\|f_a(\tilde{x}_a)\|_2 \, \|f_a(x_a)\|_2}. \tag{5}$$

This formulation is inspired by encoder-only adversarial attacks in vision-language models, which demonstrate that perturbing modality-specific encoders can induce downstream failure without relying on task supervision (Mei et al., 2025).

3.2.3 Vision Attention Suppression Loss

Attention mechanisms play a central role in multimodal models by mediating cross-modal alignment. Prior work has shown that adversarial manipulation of attention patterns can induce failures such as jailbreaking and hallucination (Wang et al., a,b). We design an objective that reduces the model's reliance on visual evidence by suppressing attention allocated to vision tokens.

Let $A_{l,h,t}(x_a + \delta, x_v, q; \theta)$ denote the attention logit at layer l, head h, and target token position t. Let $\mathcal{T}_v$ denote the set of token indices associated with the vision modality. We aggregate the total attention mass assigned to vision tokens as

$$S_v(\delta) = \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{t \in \mathcal{T}_v} A_{l,h,t}(x_a + \delta, x_v, q; \theta). \tag{6}$$

The adversarial objective is

$$\mathcal{L}^{(\text{visionatt})}(\delta) = S_v(\delta), \tag{7}$$

which is minimized during optimization to suppress visual grounding.

3.2.4 Audio Attention Amplification Loss

Complementary to suppressing visual attention, we consider an objective that explicitly amplifies the model's reliance on the perturbed audio stream. Let $\mathcal{T}_a$ denote the set of token indices corresponding to the audio modality. We aggregate the total attention mass assigned to audio tokens as

$$S_a(\delta) = \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{t \in \mathcal{T}_a} A_{l,h,t}(x_a + \delta, x_v, q; \theta). \tag{8}$$

To encourage the model to over-weight the attacked audio channel, we define

$$\mathcal{L}^{(\text{audioatt})}(\delta) = -S_a(\delta), \tag{9}$$

which is minimized during optimization.

3.2.5 Attention Randomization Loss

Beyond re-weighting attention toward or away from specific modalities, we consider an objective that directly disrupts the structure of attention itself. For each layer and head, we construct a randomized attention matrix $\tilde{A}_{l,h}$ with entries sampled uniformly within the observed range of $A_{l,h}$ and masked to preserve autoregressive structure:

$$\tilde{A}_{l,h} = \text{tril}\Big((\max(A_{l,h}) - \min(A_{l,h})) \cdot \text{Uniform}(0,1) + \min(A_{l,h}) \Big). \tag{10}$$

We then quantify the divergence using KL divergence:

$$\mathcal{L}^{(\text{randatt})}(\delta) = \sum_{l=1}^{L} \sum_{h=1}^{H} \text{KL}\Big(\text{softmax}(A_{l,h}) \,\Big\|\, \text{softmax}(\tilde{A}_{l,h}) \Big). \tag{11}$$

Minimizing this loss pushes attention toward randomized configurations.

3.2.6 Hidden-State Similarity Loss

We define an adversarial objective that targets hidden representations across transformer layers. Prior work shows that steering hidden states can induce jailbreak behaviors in large language models (Shu et al., 2025). Let $h_l(x)$ denote the hidden states produced by layer l. We define

$$L^{(\text{hidden-cos})}(\tilde{x}_a, x_a) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{|h_l|} \sum_i \cos\big(h_l^{(i)}(\tilde{x}_a), h_l^{(i)}(x_a)\big). \tag{12}$$

where $\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}$ denotes cosine similarity. Minimizing this loss pushes attacked hidden representations away from their clean directions.

3.2.7 Combined Loss

We also consider a unified attack that jointly optimizes all proposed losses:

$$\begin{aligned}\mathcal{L}^{(\text{combined})}(\delta) = \mathcal{L}_{\text{negLM}} &+ \mathcal{L}^{(\cos)}(\tilde{x}_a, x_a) \\ &+ \mathcal{L}^{(\text{visionatt})}(\delta) + \mathcal{L}^{(\text{audioatt})}(\delta) \\ &+ \mathcal{L}^{(\text{randatt})}(\delta) \\ &+ \mathcal{L}^{(\text{hidden-cos})}(\tilde{x}_a, x_a).\end{aligned} \tag{13}$$

This combined formulation aggregates complementary failure modes into a single adversarial signal. Detailed mathematical formulations for all loss functions are provided in Appendix C, and the optimization algorithm is described in Appendix D.

4 Experimental Setup

4.1 Models

The primary model used for evaluation of all six attack objectives is VideoLLAMA2 (Cheng et al., 2024). We additionally use Qwen 2.5 Omni (Xu et al., 2025a) and Qwen 3 Omni (Xu et al., 2025b) to train attacks using $\mathcal{L}_{\text{negLM}}$ and $\mathcal{L}^{(\cos)}$ and to evaluate our attacks in the black-box setting. VideoSALMONN2 (Tang et al., 2025) is also used for testing the trained attacks in the black-box settings. For Audio-Speech recognition evaluation, we use Whisper Large-v2 on LibriSpeech (Panayotov et al., 2015). Details about these models are provided in Appendix A.1.

4.2 Datasets

We evaluate on three benchmarks: AVQA (Yang et al., 2022) for multiple-choice question answering, AVSD (Alamri et al., 2019) for video-based textual summarization, and Music-AVQA (Li et al., 2022) for short answers on musical performances. Further details are provided in Appendix B. Implementation details including hyperparameters and reproducibility settings are in Appendix E.

4.3 Evaluation Metrics

Attack Success Rate We evaluate an untargeted, audio-only adversary using the Attack Success Rate (ASR), defined as the fraction of originally correct predictions that are flipped by the attack. Let $\mathcal{D} = \{(x_a^{(i)}, x_v^{(i)}, q^{(i)}, a^{\star(i)})\}_{i=1}^{N}$ be the evaluation set. For each example i, let $\tilde{x}_a^{(i)} = x_a^{(i)} + \delta^{(i)}$ denote the perturbed audio. We define

$$\text{CleanCorrect}^{(i)} := \mathbf{1}\left[\hat{a}(x_a^{(i)}, x_v^{(i)}, q^{(i)}; \theta) = a^{\star(i)}\right], \tag{14}$$

$$\text{AttackSuccess}^{(i)} := \mathbf{1}\Big[\text{CleanCorrect}^{(i)} = 1 \wedge \hat{a}(\tilde{x}_a^{(i)}, x_v^{(i)}, q^{(i)}; \theta) \neq a^{\star(i)}\Big]. \tag{15}$$

The Attack Success Rate is then

$$\text{ASR} = \frac{\sum_{i=1}^{N} \text{AttackSuccess}^{(i)}}{\sum_{i=1}^{N} \text{CleanCorrect}^{(i)}}. \tag{16}$$

Other Metrics To quantify imperceptibility of audio perturbations, we use LPIPS (Learned Perceptual Image Patch Similarity) (Zhang et al., 2018) and SI-SNR (Scale-Invariant Signal-to-Noise Ratio). LPIPS measures perceptual similarity between two signals using deep feature representations. Since LPIPS is originally defined for images, we compute it on log-mel spectrogram representations of audio. For Whisper evaluation, we use word error rate (WER). For AVSD, evaluation is performed using LLM-as-a-judge (Zheng et al., 2023). Instances where LLM judgment flips from "CORRECT" to "INCORRECT" are considered attack success. Further details about evaluation metrics are in Appendix F, G.

5 Results and Analysis

We organize our findings into three categories: (1) attack effectiveness and transferability, (2) perceptual and optimization analysis, and (3) attention and hidden-state dynamics.

5.1 Attack Effectiveness and Transferability

ENCODER-SPACE ATTACKS DOMINATE. **Attacks that directly perturb audio encoder representations consistently outperform objectives operating at the output, attention, or hidden-state levels.** As shown in Table 2, the encoder-based cosine similarity loss $\mathcal{L}^{(\cos)}$ achieves an ASR of 89.12% on AVQA, substantially exceeding attention-based and hidden-state objectives trained under identical conditions. This dominance persists when training on Music-AVQA (Table 4), where $\mathcal{L}^{(\cos)}$ attains 89.07% ASR on the source domain. These results indicate that the audio encoder constitutes a critical vulnerability bottleneck: small directional shifts in encoder embedding space are sufficient to corrupt downstream cross-modal reasoning.

Attack Objective	AVQA	Music-AVQA	AVSD
$\mathcal{L}_{\text{negLM}}$	10.27	10.74	50.43
$\mathcal{L}^{(\cos)}$	89.12	12.92	57.03
$\mathcal{L}^{(\text{visionatt})}$	18.72	7.94	49.34
$\mathcal{L}^{(\text{audioatt})}$	56.21	9.81	43.85
$\mathcal{L}^{(\text{randatt})}$	17.24	9.81	45.67
$\mathcal{L}^{(\text{hidden-cos})}$	15.13	10.28	41.53
$\mathcal{L}^{(\text{combined})}$	**96.03**	**13.80**	**59.48**

Table 2: Attack success rate (%) for different loss functions trained on VideoLLAMA2 for 150 epochs on AVQA. Evaluated on 2000 AVQA and Music-AVQA samples and the full AVSD validation set.

LIMITED CROSS-MODEL TRANSFERABILITY. **Adversarial perturbations learned on one architecture fail to transfer effectively to other multimodal models.** Table 3 shows that attacks trained on VideoLLAMA2 achieve an ASR of 10.27% on the source model but drop sharply when evaluated on Qwen 2.5 Omni (4.61%) and Qwen 3 Omni (3.62%). Similarly, perturbations optimized on Qwen models yield low ASR when applied to alternative architectures. This lack of transferability suggests that attacks exploit model-specific representational characteristics rather than universal multimodal vulnerabilities.

Source Model	VL2	Q2.5	Q3	VSAL2
VideoLLAMA2	**10.27**	4.61	3.62	5.97
Qwen 2.5 Omni	1.97	**5.10**	1.12	3.71
Qwen 3 Omni	1.80	3.80	**4.40**	3.12

Table 3: Black-box transfer results for $\mathcal{L}_{\text{negLM}}$. VL2 = VideoLLAMA2, Q2.5 = Qwen 2.5 Omni, Q3 = Qwen 3 Omni, VSAL2 = VideoSALMONN2. Bold indicates source model (white-box).

LIMITED CROSS-DOMAIN TRANSFER. **Attacks trained on one domain exhibit limited transfer to domains with different acoustic characteristics.** From Table 2, attacks trained on AVQA transfer reasonably to AVSD (59.48% for $\mathcal{L}^{(\text{combined})}$), but achieve only 13.80% ASR on Music-AVQA, a dataset focused on musical performances. Similarly, Table 4 shows that attacks

trained on Music-AVQA achieve 89.07% ASR on the source domain but only 2.30% and 24.55% on AVQA and AVSD, respectively.

Attack Objective	AVQA	Music-AVQA	AVSD
$\mathcal{L}_{\text{negLM}}$	2.00	16.60	**30.88**
$\mathcal{L}^{(\text{cos})}$	2.30	**89.07**	24.55
$\mathcal{L}^{(\text{hidden-cos})}$	2.20	**55.32**	21.93

Table 4: ASR (%) for attacks trained on Music-AVQA for 150 epochs.

ENCODER-SPACE ATTACKS DO NOT TRANSFER ACROSS ARCHITECTURES. **Encoder-space adversarial perturbations are highly specific to the audio encoder on which they are optimized.** Table 5 shows that attacks trained on the VideoLLAMA2 encoder reduce cosine similarity to 0.50 on the same encoder, but yield higher similarity when evaluated on Qwen 2.5 Omni (0.75), Qwen 3 Omni (0.70), or PANN (0.62) (Kong et al., 2020). This pattern indicates that perturbations exploit encoder-specific feature geometries rather than inducing generic acoustic distortions.

Encoder	VideoLLAMA2	Qwen 2.5 Omni	Qwen 3 Omni
VideoLLAMA2	**0.50**	0.75	0.70
Qwen 2.5 Omni	0.77	**0.74**	0.75
Qwen 3 Omni	0.96	0.93	**0.93**
PANN	0.62	0.71	0.74

Table 5: Cosine similarity across different audio encoders for $\mathcal{L}^{(\text{cos})}$ attacks. Lower similarity indicates stronger attack effect.

5.2 Perceptual and Optimization Analysis

EFFECTIVE ATTACKS CAN HAVE LOW PERCEPTUAL DISTORTION. **Highly effective attacks can be achieved with relatively small perceptual distortion to the original audio.** Table 6 and Figure 7 (see Appendix J for detailed waveform analysis) show that encoder-space and attention-based attacks such as $\mathcal{L}^{(\text{cos})}$ and $\mathcal{L}^{(\text{audioatt})}$ achieve strong attack success while maintaining low LPIPS (0.08 and 0.06) and near-zero or positive SI-SNR (-1.77 and 0.33). In contrast, $\mathcal{L}_{\text{negLM}}$ incurs substantially higher distortion (LPIPS 0.22, SI-SNR -11.48). These results demonstrate that adversarial failures can arise from subtle, structured perturbations rather than large or perceptually dominant noise.

EXTENDED OPTIMIZATION OUTPERFORMS LARGER DATASETS. **Attack effectiveness is driven more by sustained optimization than by dataset size.** Figure 2 shows that attacks trained on relatively small datasets but optimized for many iterations consistently achieve higher ASR than attacks trained on larger datasets for fewer epochs. Configurations with only a few thousand training samples but extended optimization reach ASR values exceeding 80%, whereas attacks trained on tens of thousands of samples with limited iterations remain below 20% ASR. This indicates that adversarial optimization benefits from repeatedly exploiting consistent model behaviors present in a small set of inputs. See App. H for exact values.

Attack Objective	LPIPS (↓)	SI-SNR (dB) (↑)
$\mathcal{L}_{\text{negLM}}$	0.22	-11.48
$\mathcal{L}^{(\text{cos})}$	0.08	-1.77
$\mathcal{L}^{(\text{visionatt})}$	0.07	-1.02
$\mathcal{L}^{(\text{audioatt})}$	0.06	0.33
$\mathcal{L}^{(\text{randatt})}$	0.06	0.05
$\mathcal{L}^{(\text{hidden-cos})}$	0.06	0.56
$\mathcal{L}^{(\text{combined})}$	0.14	-6.23

Table 6: LPIPS of log-mel spectrograms and SI-SNR for each loss trained on VideoLLAMA2. Lower LPIPS and higher SI-SNR indicate less perceptible perturbations.

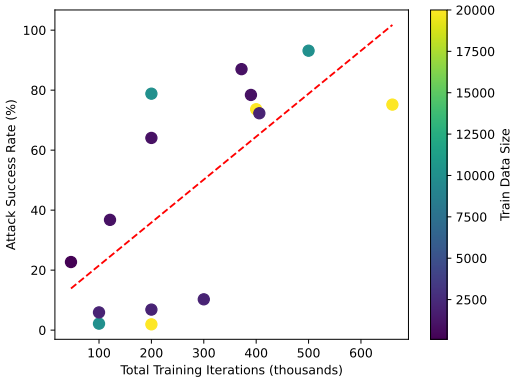

Figure 2: Relation between total training iterations and ASR for $\mathcal{L}_{\text{negLM}}$. Each point corresponds to an attack configuration, colored by training data size. Extended optimization on smaller datasets yields higher ASR.

NON-MONOTONIC ATTACK BUDGET EFFECTS. **Increasing the perturbation budget does not monotonically improve attack effectiveness.** For attack budget analysis, each attack is trained until it converges (conditions for convergence are in Appendix E.3). Figure 3 shows that ASR improves from 49.2% at budget 0.3 to 73.51% at budget 1.0, even though both settings converge in a similar number of epochs (213 vs. 209). In contrast,

a mid-range budget of 0.7 underperforms substantially (51.89% ASR) despite requiring more epochs to converge than 1.0 (265 vs. 209). This pattern suggests that the attack constraint interact with optimization dynamics in complex ways: some budgets admit highly effective perturbations that are also easier to optimize (budget 1.0), while others converge to weaker solutions even with extended training (budget 0.7).

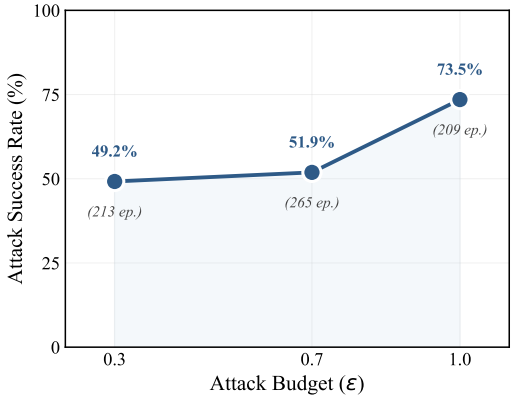

Figure 3: Attack budget vs ASR for $\mathcal{L}^{(\cos)}$ on VideoLLAMA2. Each attack was trained until convergence.

PHYSICAL-WORLD ROBUSTNESS VIA RIR AUGMENTATION. **Training with room impulse response (RIR) improves robustness but physical attacks remain weaker than digital ones.** To evaluate the feasibility of our attack under realistic acoustic conditions, we incorporate a room impulse response (RIR) regularizer during training, which simulates environmental effects such as reverberation and signal propagation. We then compare attack success rates for digitally applied and physically simulated (RIR) perturbation for the same attack (Table 7).

Objectives	Digitally Applied Attack	Physical Attack (RIR)
$\mathcal{L}_{\text{negLM}}$	82.32	39.27
$\mathcal{L}^{(\cos)}$	81.18	22.22
$\mathcal{L}^{(\text{visionatt})}$	39.44	32.32
$\mathcal{L}^{(\text{audioatt})}$	82.58	98.12
$\mathcal{L}^{(\text{randatt})}$	35.41	20.80
$\mathcal{L}^{(\text{hidden-cos})}$	77.04	87.34
$\mathcal{L}^{(\text{combined})}$	**83.68**	**97.96**

Table 7: Attack success rate (%) under digital and physically simulated (RIR) settings.

In most cases, physical attacks exhibit lower success rates than digital ones. For example, $\mathcal{L}_{\text{negLM}}$ drops from 82.32% to 39.27%, and $\mathcal{L}^{(\cos)}$ decreases from 81.18% to 22.22%. Similarly, $\mathcal{L}^{(\text{randatt})}$ reduces from 35.41% to 20.80%, indicating that environmental distortions degrade adversarial effectiveness. We hypothesize that this reduction arises from the sensitivity of audio signals to environmental factors, where the recorded waveform depends on room geometry, materials, and microphone characteristics. However, certain objectives remain robust under physical transformations. $\mathcal{L}^{(\text{audioatt})}$ increases from 82.58% to 98.12%, and $\mathcal{L}^{(\text{hidden-cos})}$ improves from 77.04% to 87.34%, suggesting that some attack mechanisms are inherently resilient to acoustic variability. Notably, the combined objective achieves higher performance in the physical setting (97.96%) compared to digital (83.68%), indicating that RIR-based training can enhance robustness.

These results demonstrate that while physical deployment introduces variability that weakens several objectives, RIR-augmented training enables attacks to remain effective in realistic acoustic environments. Further discussion on the realism of the threat model and broader implications is provided in Appendix K.

5.3 Attention, Hidden-State and Output Dynamics

ATTACK SUCCESS ARISES FROM DISTORTIONS, NOT LOW BASELINE. **High ASR is driven by attack-induced distortions rather than low baseline model performance.** Table 16 shows that the model maintains strong clean performance prior to attack (AVQA: 0.956, Music-AVQA: 0.807). Under adversarial perturbation, accuracy drops sharply for effective attacks: $\mathcal{L}^{(\cos)}$ reduces AVQA accuracy from 0.956 to 0.11, while the combined attack causes degradation to 0.039. In contrast, weaker objectives such as vision-attention manipulation produce only marginal accuracy changes (0.036 drop).

ADVERSARIAL RESPONSES MAINTAIN HIGH CONFIDENCE. **Adversarially induced responses are generated with confidence levels similar to clean inputs.** Table 15 shows that under the combined attack, the model's confidence on AVQA remains close between clean and adversarial inputs (0.93 vs. 0.85), despite a substantial ASR of 96.03%. Several attacks exhibit small or even negative confidence differences, indicating that adversarial failures arise from internal misalignment rather than uncertainty.

LAYER-SPECIFIC VULNERABILITIES. **The effectiveness of attention and hidden-state attacks**

depends strongly on which layers are perturbed. Table 8 shows that lower layers (1–10) contribute most to audio-driven attacks, with audio-attention and hidden-state losses achieving ASRs of 39.75% and 32.85%. Mid-level layers (11–18) are most influential for video-attention manipulation (27.92% ASR). Higher layers (19–28) yield consistently low ASR across all attack types. When all layers are jointly optimized, attack success increases substantially, confirming that different layer subsets capture complementary vulnerabilities. See Appendix I for layer-wise attention visualizations.

Layers	Audio Attention	Video Attention	Hidden
1–10	39.75	2.25	**32.85**
11–18	11.98	**27.92**	2.90
19–28	2.04	2.20	2.62
All	**56.21**	18.72	15.13

Table 8: ASR (%) by layer subset. VL2 has 28 layers.

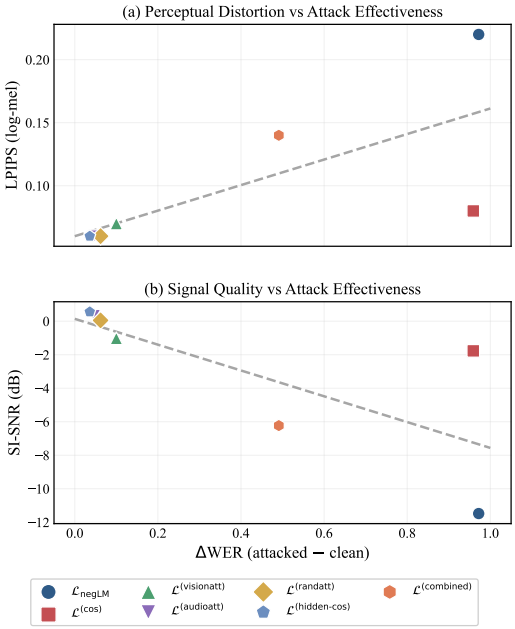

Figure 4: Re lationship between attack effectiveness on Whisper and perceptual distortion. Top: WER difference vs LPIPS. Bottom: WER difference vs SI-SNR.

AUTOMATIC SPEECH RECOGNITION SYSTEMS RESPOND TO DISTORTION MAGNITUDE. **Speech recognition models are primarily sensitive to perturbation magnitude rather than specific attack objectives.** Figure 4 and Table 13 show that larger WER increases correlate strongly with higher perceptual distortion (LPIPS). Losses that induce substantial distortion, such as $\mathcal{L}_{\text{negLM}}$ and $\mathcal{L}^{(\text{cos})}$, achieve high ASR on Whisper (98.20% and 97.38%). In contrast, attention and hidden-state attacks produce minimal distortion and correspondingly low ASR. This indicates that automatic speech recognition systems are more sensitive to overall acoustic corruption than to structured adversarial manipulations. Table 9 further quantifies this trend: objectives that induce larger distortion, such as $\mathcal{L}_{\text{negLM}}$ and $\mathcal{L}^{(\text{cos})}$, achieve near-complete transfer to Whisper (98.20% and 97.38% ASR) with large ΔWER (0.972 and 0.959). In contrast, low-distortion objectives such as $\mathcal{L}^{(\text{hidden-cos})}$ and $\mathcal{L}^{(\text{audioatt})}$ yield minimal degradation (7.37% and 9.25% ASR with ΔWER below 0.05). This reinforces that attack success is driven by distortion magnitude, not the attack objective.

Attack Objective	ASR% on Whisper	ΔWER
$\mathcal{L}_{\text{negLM}}$	**98.20**	0.972
$\mathcal{L}^{(\text{cos})}$	97.38	0.959
$\mathcal{L}^{(\text{visionatt})}$	17.99	0.100
$\mathcal{L}^{(\text{audioatt})}$	9.25	0.047
$\mathcal{L}^{(\text{randatt})}$	11.29	0.062
$\mathcal{L}^{(\text{hidden-cos})}$	7.37	0.036
$\mathcal{L}^{(\text{combined})}$	66.88	0.491

Table 9: Black-box transfer to Whisper Large-v2. ΔWER = WER(attacked) − WER(clean) on LibriSpeech. Attacks with higher perceptual distortion achieve higher ASR, indicating Whisper responds primarily to distortion magnitude.

6 Conclusion

We present SOUNDBREAK, a systematic study of audio-only adversarial attacks on trimodal audio-video-language models. Our findings reveal that carefully crafted audio perturbations can reliably induce multimodal failures without modifying visual or textual inputs. Encoder-space attacks constitute a critical vulnerability bottleneck, achieving up to 96% attack success rate. We show that these attacks exhibit limited transferability across model architectures and encoder designs, while speech recognition models respond primarily to distortion magnitude rather than structured adversarial objectives. Our analysis of layer-specific vulnerabilities, optimization dynamics, and confidence patterns provides practical insights for both attack construction and defensive design. Future work should explore adaptive defenses that monitor cross-modal consistency and investigate attack persistence under real-world conditions such as over-the-air propagation and environmental interference.

Acknowledgments

We acknowledge Advanced Research Computing at Virginia Tech for providing computational resources and technical support that have contributed to the results reported within this paper. Additionally, we also thank the Commonwealth Cyber Initiative for their support. Lastly, we thank all reviewers, area chair, and senior area chair for their comments which helped improve the paper.

Limitations

Our study has several limitations that warrant discussion. **First,** we focus exclusively on white-box attacks where the adversary has full access to model gradients and internal representations. Real-world adversaries may have limited or no access to model internals, and developing effective black-box attacks remains an open challenge. While we evaluate black-box transferability across models, the low transfer rates suggest that practical deployment would require model-specific optimization. **Second,** our evaluation is limited to four multimodal models and three benchmarks. Generalization to other architectures, particularly those with fundamentally different audio encoding strategies or cross-modal fusion mechanisms, cannot be assumed without additional experiments. **Third,** we do not evaluate potential defenses against our attacks. Promising directions include input preprocessing, adversarial training, and cross-modal consistency checking, but their efficacy and computational costs remain unexplored.

Ethics Statement

This work studies adversarial vulnerabilities in multimodal models to inform the development of more robust systems. We do not provide tools or code that could be directly used for malicious attacks. Our evaluation uses publicly available models and datasets, requiring no collection of personal data or human annotations. All models are evaluated in accordance with their respective licensing agreements. We believe that understanding these vulnerabilities is essential for deploying multimodal systems safely, and that responsible disclosure of attack surfaces enables the research community to develop appropriate defenses.

AI Assistance. We used ChatGPT for assistance with portions of the manuscript, including generating LaTeX code for tables and algorithms, and refining text drafted by the authors. All AI-generated content was carefully reviewed and revised by the authors to ensure accuracy, clarity, and consistency with our experimental findings.

References

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. 2023. Abusing images and sounds for indirect instruction injection in multi-modal llms. *Preprint*, arXiv:2307.10490.

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

Luis Felipe Chary and Miguel Arjona Ramirez. 2025. Spectrogram patch codec: A 2d block-quantized vq-vae and hifi-gan for neural speech coding. *Preprint*, arXiv:2509.02244.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Hyeongjun Choi, Ji Hyuk Jung, and Ji Won Yoon. 2024. Ghost in the radio: An audio adversarial attack using environmental noise through radio. *IEEE Access*.

Amirbek Djanibekov, Nurdaulet Mukhituly, Kentaro Inui, Hanan Aldarmaki, and Nils Lukas. 2025. Spirit: Patching speech language models against jailbreak attacks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14514–14531.

Simon Geisler, Tom Wollschläger, MHI Abdalla, Johannes Gasteiger, and Stephan Günnemann. 2024. Attacking large language models with projected gradient descent. In *ICML 2024 Next Generation of AI Safety Workshop*.

Simon Geisler, Tom Wollschläger, M. H. I. Abdalla, Johannes Gasteiger, and Stephan Günnemann. 2025. Attacking large language models with projected gradient descent. *Preprint*, arXiv:2402.09154.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *stat*, 1050:20.

Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tianwei Zhang. 2024. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 3385–3403. IEEE.

Shuyang Hao, Bryan Hooi, Jun Liu, Kai-Wei Chang, Zi Huang, and Yujun Cai. 2025. Exploring visual vulnerabilities via multi-loss adversarial search for jailbreaking vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19890–19899.

Chengze Jiang, Zhuangzhuang Wang, Minjing Dong, and Jie Gui. 2025. Survey of adversarial robustness in multimodal large language models. *arXiv preprint arXiv:2503.13962*.

Hangoo Kang, Jehyeok Yeon, and Gagandeep Singh. 2025a. Trap: Targeted redirecting of agentic preferences. *Preprint*, arXiv:2505.23518.

Mintong Kang, Chejian Xu, and Bo Li. 2025b. Advwave: Stealthy adversarial jailbreak attack against large audio-language models. In *The Thirteenth International Conference on Learning Representations*.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19108–19118.

Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. 2025. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, pages 1–20.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. 2024. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24645–24654.

Chang Liu, Haolin Wu, Xi Yang, Kui Zhang, Cong Wu, Weiming Zhang, Nenghai Yu, Tianwei Zhang, Qing Guo, and Jie Zhang. 2025a. Exploiting vulnerabilities in speech translation systems through targeted adversarial attacks. *arXiv preprint arXiv:2503.00957*.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. 2025b. A survey of attacks on large vision–language models: Resources, advances, and future trends. *IEEE Transactions on Neural Networks and Learning Systems*.

Xuannan Liu, Xing Cui, Peipei Li, Zekun Li, Huaibo Huang, Shuhan Xia, Miaoxuan Zhang, Yueying Zou, and Ran He. 2024. Jailbreak attacks and defenses against multimodal generative models: A survey. *arXiv preprint arXiv:2411.09259*.

Yuchen Liu, Apu Kapadia, and Donald Williamson. 2023. Privacy-preserving and privacy-attacking approaches for speech and audio–a survey. *arXiv preprint arXiv:2309.15087*.

Liming Lu, Shuchao Pang, Siyuan Liang, Haotian Zhu, Xiyu Zeng, Aishan Liu, Yunhuai Liu, and Yongbin Zhou. 2025. Adversarial training for multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2503.04833*.

Rao Ma, Mengjie Qian, Vyas Raina, Mark Gales, and Kate Knill. 2025. Universal acoustic adversarial attacks for flexible control of speech-llms. Association for Computational Linguistics (ACL).

Hefei Mei, Zirui Wang, Shen You, Minjing Dong, and Chang Xu. 2025. Veattack: Downstream-agnostic vision encoder attack against large vision language models. *arXiv preprint arXiv:2505.17440*.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Sahid Hossain Mustakim, SM Islam, Ummay Maria Muna, Montasir Chowdhury, Mohammed Jawwadul Islam, Sadia Ahmmed, Tashfia Sikder, Syed Tasdid Azam Dhrubo, and Swakkhar Shatabda. 2025. Watch, listen, understand, mislead: Tri-modal adversarial attacks on short videos for content appropriateness evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2955–2964.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE.

Vyas Raina, Rao Ma, Charles McGhee, Kate Knill, and Mark Gales. 2024. Muting whisper: A universal acoustic adversarial attack on speech foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7549–7565.

Vinu Sankar Sadasivan, Soheil Feizi, Rajiv Mathews, and Lun Wang. 2026. Attacker's noise can manipulate your audio-based llm in the real world. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1430–1440.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*.

Huizhen Shu, Xuying Li, Qirui Wang, Yuji Kosuga, Mengqiu Tian, and Zhuo Li. 2025. Layer-wise perturbations via sparse autoencoders for adversarial text generation. *Preprint*, arXiv:2508.10404.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*.

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

Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. 2022. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 15375–15385.

Le Wang, Zonghao Ying, Tianyuan Zhang, Siyuan Liang, Shengshan Hu, Mingchuan Zhang, Aishan Liu, and Xianglong Liu. 2025. Manipulating multimodal agents via cross-modal prompt injection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10955–10964.

Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. 2024. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6920–6928.

Xiaosen Wang, Shaokang Wang, Zhijin Ge, Yuyang Luo, and Shudong Zhang. a. Attention! your vision language model could be maliciously manipulated. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zijun Wang, Haoqin Tu, Jieru Mei, Bingchen Zhao, Yisen Wang, and Cihang Xie. b. Attngcg: Enhancing jailbreaking attacks on llms with attention manipulation. *Transactions on Machine Learning Research*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. 2024. Shadowcast: Stealthy data poisoning attacks against vision-language models. *Advances in Neural Information Processing Systems*, 37:57733–57764.

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

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. Jailbreak attacks and defenses against large language models: A survey. *Preprint*, arXiv:2407.04295.

Haiyang Yu, Tian Xie, Jiaping Gui, Pengyang Wang, Pengzhou Cheng, Ping Yi, and Yue Wu. 2025. Backdoormbti: A backdoor learning multimodal benchmark tool kit for backdoor defense evaluation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2791–2802.

Haiqi Zhang, Hao Tang, Yanpeng Sun, Shengfeng He, and Zechao Li. 2025a. Modality-specific interactive attack for vision-language pre-training models. *IEEE Transactions on Information Forensics and Security*.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. 2025b. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19900–19909.

Jiaming Zhang, Qi Yi, and Jitao Sang. 2022. Towards adversarial attack on vision-language pre-training

models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, Zhanhui Kang, and Yu Wang. 2025c. Qava: Query-agnostic visual attack to large vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10205–10218.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Contents of the Appendix

A Experimental Setup Details

A.1 Model Architectures

We evaluate on three multimodal models and two audio-only systems. Table 10 summarizes the key architectural differences.

Model	Layers	Audio Encoder	LM Backbone
VideoLLAMA2	28	BEATs	Qwen2-7B-Instruct
Qwen 2.5 Omni	28	Whisper-style	Qwen2.5
Qwen 3 Omni	28	Whisper-style	Qwen3
video SALMONN 2	36	Whisper	Qwen2.5-VL 3B based
Whisper Large-v2	32	-	-
PANNs	14	CNN	N/A

Table 10: Summary of model architectures used in our experiments.

VideoLLAMA2. VideoLLAMA2 (Cheng et al., 2024) serves as our primary evaluation model. The architecture uses a CLIP-based visual encoder and a separate audio encoder that extracts 128-dimensional mel-filterbank features. Audio embeddings are projected into the language model space and concatenated with visual tokens before processing by a Mistral-7B backbone. The model has 28 transformer layers, which we partition into early (1–10), middle (11–18), and late (19–28) subsets for layer-wise analysis. The checkpoint used for VideoLLAMA2 is "DAMO-NLP-SG/VideoLLaMA2.1-7B-AV".

Qwen 2.5 Omni and Qwen 3 Omni. Qwen 2.5 Omni (Xu et al., 2025a) and Qwen 3 Omni (Xu et al., 2025b) use a Thinker-Talker architecture where the Thinker processes multimodal inputs and the Talker generates text or speech outputs. Both models use a Whisper-style audio encoder with different tokenization than VideoLLAMA2. We use these models to evaluate whether attacks transfer across different audio processing pipelines. Qwen 3 Omni is a more recent release with updated training, providing a test case for transfer to newer model versions. The checkpoints used

for Qwen 2.5 Omni and and Qwen 3 Omni are Qwen/Qwen2.5-Omni-7B and Qwen/Qwen3-30B-A3B-Instruct-2507.

video-SALMONN 2. video-SALMONN 2 (Tang et al., 2025) is an audio-visual large language model designed for detailed video captioning and question answering. The architecture processes audio and visual streams using separate encoders, followed by modality-specific aligners that project features into a shared language model space. Audio tokens and visual tokens are temporally aligned and interleaved before being combined with text tokens and passed to the LLM backbone for generation. The model uses a Whisper-based audio encoder and a visual encoder derived from a pretrained visual LLM, while keeping the backbone frozen during training. We use video-SALMONN 2 to evaluate the transferability of our attack to different architectures. The checkpoint used is tsinghua-ee/video-SALMONN2_plus_3B_full.

Whisper Large-v2. Whisper is an encoder-decoder speech recognition model trained on 680,000 hours of web audio. The encoder processes 80-dimensional log-mel spectrograms; the decoder generates transcriptions autoregressively. We use Whisper to evaluate black-box transfer of attacks to speech-only systems outside the trimodal setting.

PANNs. Pretrained Audio Neural Networks (PANNs) (Kong et al., 2020) are CNNs trained on AudioSet for general audio tagging. We use PANNs as an external encoder to test whether $\mathcal{L}^{(\cos)}$ attacks transfer across encoder architectures or exploit VideoLLAMA2-specific features.

B Dataset Descriptions

This section provides detailed descriptions of the datasets used for training and evaluating our adversarial attacks. We select datasets that span different audio-visual reasoning tasks to comprehensively assess attack effectiveness and transferability.

AVQA (Audio-Visual Question Answering). AVQA (Yang et al., 2022) is a large-scale audio-visual question answering dataset introduced at ACM MM 2022. The dataset contains approximately 57,015 videos sourced from VGG-Sound, depicting real audio-visual scenes with natural sounds and objects. Each video is associated with one or more question-answer pairs, totaling 57,335 QA instances. The questions are designed such that answering them often requires integrating both audio and visual cues; many questions cannot be answered correctly using a single modality alone. This cross-modal dependency makes AVQA particularly suitable for evaluating whether audio-only perturbations can disrupt multimodal reasoning even when visual information remains unperturbed. We use the original train/validation/test splits provided by the dataset authors. In our experiments, AVQA serves as the primary benchmark for both attack training and evaluation.

Music-AVQA. Music-AVQA (Li et al., 2022) is a short-answer audio-visual question answering dataset focused on musical performance videos. The dataset requires fine-grained spatio-temporal reasoning across modalities, with questions that probe understanding of musical instruments, performer actions, and acoustic properties. Videos feature diverse musical performances including solo instruments, ensembles, and various musical genres. The acoustic characteristics of musical content differ substantially from the speech and environmental sounds that dominate AVQA, providing a challenging test case for cross-domain transfer. We use Music-AVQA to evaluate whether attacks trained on general audio-visual scenarios generalize to structured music-centric scenes with distinct spectral properties.

AVSD (Audio-Visual Scene-Aware Dialog). AVSD (Alamri et al., 2019) is a dialog-based dataset from DSTC7/DSTC8 where models generate free-form, natural language responses grounded in video content, audio information, and dialog history. The dataset is built over more than 11,000 videos from the Charades dataset, which depicts people performing everyday activities in home environments. Each sample includes a 10-round dialog history plus a final summary describing the video content. Unlike AVQA and Music-AVQA which use short-answer formats, AVSD requires generating extended natural language descriptions, providing a complementary evaluation of attack effectiveness on generative tasks. We use AVSD to assess cross-dataset transferability within the same domain, particularly for evaluating whether attacks transfer across different output modalities (multiple-choice versus open-ended generation).

LibriSpeech. LibriSpeech (Panayotov et al., 2015) is a large-scale corpus of read English speech

derived from audiobooks in the LibriVox project. The dataset contains approximately 1,000 hours of speech at 16kHz sampling rate, with high-quality transcriptions suitable for training and evaluating automatic speech recognition systems. The dataset is partitioned into training, development, and test sets with varying levels of difficulty based on speaker characteristics and recording conditions. In our experiments, LibriSpeech is used solely for evaluating black-box transfer of adversarial perturbations to Whisper, enabling controlled analysis of attack effectiveness in a speech-only setting without the confounding factors of visual input or multimodal reasoning.

C Mathematical Formulation of Attack Objectives

This section provides detailed mathematical formulations for each adversarial objective used in our attack framework. We describe the theoretical motivation, precise mathematical definitions, and implementation considerations for each loss function.

C.1 Perturbation Model and Constraint Set

Let $x_a \in \mathbb{R}^T$ denote the clean audio waveform of length T samples at sampling rate $sr = 16\text{kHz}$. The adversary constructs a perturbation $\delta \in \mathbb{R}^{T_\delta}$ of fixed length $T_\delta = 10 \cdot f_s = 160{,}000$ samples, corresponding to a 10-second audio segment. For audio inputs longer than T_δ, the perturbation is repeated cyclically:

$$\delta_{\text{extended}}[t] = \delta[t \mod T_\delta], \quad t = 0, 1, \ldots, T-1. \tag{17}$$

The perturbed audio is then computed as the element-wise sum:

$$\tilde{x}_a = x_a + \cdot\delta_{\text{extended}}[:T], \tag{18}$$

The notation $[:T]$ denotes truncation to the first T elements.

To ensure bounded perturbation magnitude, we constrain δ to lie within an ℓ_∞ ball:

$$\mathcal{C}_\varepsilon = \{\delta \in \mathbb{R}^{T_\delta} : \|\delta\|_\infty \leq \varepsilon\}, \tag{19}$$

where $\varepsilon > 0$ is the attack budget. After each gradient update, the perturbation is projected back onto $\mathcal{C}_\varepsilon$ via element-wise clipping:

$$\Pi_{\mathcal{C}_\varepsilon}(\delta) = \text{clip}(\delta, -\varepsilon, \varepsilon). \tag{20}$$

Following perturbation application, the combined waveform is normalized to prevent amplitude overflow. We employ a normalization scheme that preserves the relative contribution of the perturbation:

$$\tilde{x}_a^{\text{norm}} = \frac{\tilde{x}_a}{\max(|\tilde{x}_a|) + \epsilon}, \tag{21}$$

where $\epsilon = 10^{-8}$ is a small constant for numerical stability. This normalization ensures that the final audio remains within the valid amplitude range $[-1, 1]$ while maintaining the adversarial signal structure.

C.2 Negative Language Modeling Loss

The negative language modeling loss directly targets the model's output distribution by minimizing the probability assigned to the correct answer. Let $a^\star = (a_1^\star, a_2^\star, \ldots, a_m^\star)$ denote the ground-truth answer sequence of m tokens. The standard cross-entropy language modeling loss under perturbed audio is:

$$\mathcal{L}_{\text{LM}}(\delta) = -\sum_{j=1}^{m} \log p_\theta(a_j^\star \mid a_{<j}^\star, \tilde{x}_a, x_v, q), \tag{22}$$

where p_θ denotes the model's conditional probability distribution, $a_{<j}^\star$ represents the preceding tokens, x_v is the video input, and q is the question.

To suppress the correct answer, we negate this loss:

$$\begin{aligned}\mathcal{L}_{\text{negLM}}(\delta) &= -\mathcal{L}_{\text{LM}}(\delta) \\ &= \sum_{j=1}^{m} \log p_\theta\big(a_j^\star \mid a_{<j}^\star, \tilde{x}_a, x_v, q\big)\,.\end{aligned} \tag{23}$$

Minimizing $\mathcal{L}_{\text{negLM}}$ is equivalent to minimizing the log-probability of the correct answer sequence. Intuitively, this objective pushes the model's probability mass away from the correct answer toward alternative responses.

This formulation connects to recent work on attacking large language models through projected gradient descent on input representations (Geisler et al., 2025). Unlike token-space attacks that require discrete optimization, our approach operates in the continuous audio waveform space where standard gradient methods apply directly.

C.3 Encoder-Based Cosine Similarity Loss

The encoder-based attack targets the audio representation space rather than the output distribution. Let $f_a : \mathbb{R}^{N_f \times 128} \to \mathbb{R}^{N_e \times d}$ denote the

audio encoder that maps filterbank features to a sequence of N_e embeddings of dimension d. We further apply a modality-specific projection layer $g_a : \mathbb{R}^{N_e \times d} \to \mathbb{R}^{N_e \times d_{\text{lm}}}$ that aligns audio embeddings with the language model's hidden dimension d_{lm}.

For clean audio x_a with filterbank features $\mathbf{F}$ and perturbed audio $\tilde{x}_a$ with features $\tilde{\mathbf{F}}$, we compute the projected embeddings:

$$\mathbf{E}_{\text{clean}} = g_a(f_a(\mathbf{F})) \in \mathbb{R}^{N_e \times d_{\text{lm}}}, \tag{24}$$

$$\mathbf{E}_{\text{adv}} = g_a(f_a(\tilde{\mathbf{F}})) \in \mathbb{R}^{N_e \times d_{\text{lm}}}. \tag{25}$$

The cosine similarity loss aggregates pairwise similarities across all embedding positions:

$$\mathcal{L}^{(\cos)}(\delta) = \frac{1}{N_e} \sum_{i=1}^{N_e} \frac{\mathbf{E}_{\text{clean}}[i] \cdot \mathbf{E}_{\text{adv}}[i]}{\|\mathbf{E}_{\text{clean}}[i]\|_2 \cdot \|\mathbf{E}_{\text{adv}}[i]\|_2}, \tag{26}$$

where $\mathbf{E}[i] \in \mathbb{R}^{d_{\text{lm}}}$ denotes the i-th embedding vector. Minimizing this loss pushes the adversarial embeddings to be orthogonal to their clean counterparts in the projected space.

The motivation for targeting the encoder rather than the output stems from the observation that encoder representations serve as the foundation for all downstream processing. By corrupting the audio representation at this early stage, we can induce failures that propagate through subsequent attention and decoding mechanisms without requiring supervision from specific output targets.

C.4 Vision Attention Suppression Loss

Attention mechanisms in transformer models determine how information flows between different input components. In multimodal models, cross-modal attention allows the model to ground its reasoning in relevant modalities. We design an objective that reduces the model's reliance on visual information by suppressing attention to vision tokens.

Let $\mathbf{A}^{(l,h)} \in \mathbb{R}^{N \times N}$ denote the attention matrix at layer $l \in \{1, \ldots, L\}$ and head $h \in \{1, \ldots, H\}$, where N is the total sequence length including all modality tokens. Let $\mathcal{T}_v = \{v_{\text{start}}, v_{\text{start}} + 1, \ldots, v_{\text{end}}\}$ denote the indices of vision tokens in the input sequence.

The total attention mass assigned to vision tokens is:

$$S_v(\delta) = \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{N} \sum_{j \in \mathcal{T}_v} \mathbf{A}_{i,j}^{(l,h)}, \tag{27}$$

where $\mathbf{A}_{i,j}^{(l,h)}$ represents the attention weight from position i to position j. The vision attention suppression loss is simply:

$$\mathcal{L}^{(\text{visionatt})}(\delta) = S_v(\delta). \tag{28}$$

Minimizing this loss reduces the total attention directed toward visual tokens across all layers and heads, effectively "blinding" the model to visual information even though the video input remains unchanged.

C.5 Audio Attention Amplification Loss

Complementary to suppressing visual attention, we design an objective that amplifies attention to audio tokens. The intuition is that by forcing the model to over-attend to the perturbed audio stream, we can maximize the influence of adversarial information on the model's reasoning process.

Let $\mathcal{T}_a = \{a_{\text{start}}, a_{\text{start}} + 1, \ldots, a_{\text{end}}\}$ denote the indices of audio tokens. The total attention mass on audio tokens is:

$$S_a(\delta) = \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{N} \sum_{j \in \mathcal{T}_a} \mathbf{A}_{i,j}^{(l,h)}. \tag{29}$$

To maximize this quantity, we define the loss as its negation:

$$\mathcal{L}^{(\text{audioatt})}(\delta) = -S_a(\delta). \tag{30}$$

Minimizing $\mathcal{L}^{(\text{audioatt})}$ is equivalent to maximizing S_a, thereby increasing the model's reliance on audio tokens.

This objective is particularly effective when combined with encoder-space attacks. By simultaneously corrupting the audio representation and forcing the model to attend to it, we create a compounding effect where the model not only receives corrupted information but is also compelled to rely on it for reasoning.

C.6 Attention Randomization Loss

Beyond re-weighting attention across modalities, we consider an objective that directly disrupts the structure of attention patterns. The goal is to push attention matrices toward random configurations that lack the meaningful structure learned during training.

For each attention matrix $\mathbf{A}^{(l,h)}$, we construct a randomized target $\tilde{\mathbf{A}}^{(l,h)}$ as follows. Let $a_{\max} = \max(\mathbf{A}^{(l,h)})$ and $a_{\min} = \min(\mathbf{A}^{(l,h)})$ denote the

range of attention logits. We sample a random matrix:

$$\mathbf{R}^{(l,h)} = (a_{\max} - a_{\min}) \cdot \mathbf{U} + a_{\min}, \tag{31}$$

where $\mathbf{U} \in \mathbb{R}^{N \times N}$ has entries sampled uniformly from $[0, 1]$. To preserve the autoregressive structure of causal attention, we apply a lower-triangular mask:

$$\tilde{\mathbf{A}}^{(l,h)} = \mathbf{R}^{(l,h)} \odot \mathbf{M}_{\text{tril}}, \tag{32}$$

where $\mathbf{M}_{\text{tril}}$ is a lower-triangular matrix of ones and $\odot$ denotes element-wise multiplication.

The attention randomization loss measures the divergence between actual and randomized attention using KL divergence over the softmax-normalized distributions:

$$\mathcal{L}^{(\text{randatt})}(\delta) = \sum_{l=1}^{L} \sum_{h=1}^{H} D_{\text{KL}}\Big(\text{softmax}(\mathbf{A}^{(l,h)}) \,\Big\|\, \text{softmax}(\tilde{\mathbf{A}}^{(l,h)})\Big), \tag{33}$$

where $D_{\text{KL}}(P\|Q) = \sum_i P_i \log(P_i/Q_i)$ is the Kullback-Leibler divergence and the softmax is applied row-wise.

Minimizing this loss pushes the attention patterns toward the randomized targets. Since the targets are sampled independently at each forward pass, this objective effectively encourages attention patterns that are uniformly random rather than structured around semantically meaningful relationships.

C.7 Hidden-State Similarity Loss

The hidden-state attack targets the internal representations at each transformer layer. Unlike the encoder-based attack which operates only on the audio encoder outputs, this objective propagates through the entire model architecture.

Let $\mathbf{h}_l(x) \in \mathbb{R}^{N \times d_{\text{lm}}}$ denote the hidden states at layer l for input x. For clean input (x_a, x_v, q) and perturbed input $(\tilde{x}_a, x_v, q)$, we compute hidden states at each layer and measure their similarity using cosine distance:

$$\mathcal{L}^{(\text{hidden-cos})}(\delta) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{N} \sum_{i=1}^{N} \cos\Big(\mathbf{h}_l^{(i)}(\tilde{x}_a),\, \mathbf{h}_l^{(i)}(x_a)\Big), \tag{34}$$

where $\mathbf{h}_l^{(i)} \in \mathbb{R}^{d_{\text{lm}}}$ denotes the hidden state at position i in layer l.

This formulation requires computing a full forward pass for both clean and perturbed inputs, then comparing their hidden representations position-by-position and layer-by-layer. Minimizing this loss pushes the adversarial hidden states to diverge from their clean counterparts throughout the network depth.

The hidden-state attack is motivated by work on representation engineering and activation steering in language models (Shu et al., 2025), which demonstrates that manipulating internal representations can induce specific behavioral changes. By targeting hidden states, we can potentially induce more fundamental failures in the model's reasoning process compared to output-level attacks.

C.8 Combined Loss

To leverage the complementary failure modes induced by different attack objectives, we combine all losses into a unified objective:

$$\begin{aligned}\mathcal{L}^{(\text{combined})}(\delta) = \mathcal{L}_{\text{negLM}}(\delta) + \mathcal{L}^{(\text{cos})}(\delta) \\ + \mathcal{L}^{(\text{visionatt})}(\delta) + \mathcal{L}^{(\text{audioatt})}(\delta) \\ + \mathcal{L}^{(\text{randatt})}(\delta) + \mathcal{L}^{(\text{hidden-cos})}(\delta).\end{aligned} \tag{35}$$

All loss components are combined with equal weighting. We do not tune individual loss weights, as our goal is to assess whether a simple combination of diverse objectives can outperform individual attacks.

The combined loss aggregates signals from multiple stages of the audio-to-output pipeline: encoder representations, cross-modal attention allocation, internal hidden states, and output likelihood. This multi-view formulation provides gradient information that can guide the perturbation toward configurations that simultaneously disrupt multiple computational mechanisms.

D Optimization Algorithm

This section presents the complete optimization procedure for learning adversarial audio perturbations. We decompose the attack into three modular algorithms: audio preprocessing (Algorithm 1), single-step gradient update (Algorithm 2), and the main optimization loop (Algorithm 3).

D.1 Audio Preprocessing

Algorithm 1 describes the audio preprocessing pipeline that transforms raw waveforms into model-ready features. This procedure handles perturba-

tion application, amplitude normalization, and mel-filterbank extraction.

Algorithm 1 Audio Preprocessing Pipeline

Require: Clean audio $x_a \in \mathbb{R}^T$, perturbation $\delta \in \mathbb{R}^{T_\delta}$
Ensure: Mel-filterbank features $\mathbf{F}$

// Cyclic perturbation extension
for $t = 0$ to $T - 1$ **do**
  $\tilde{x}_a[t] \leftarrow x_a[t] + \delta[t \mod T_\delta]$
end for
// Amplitude normalization
$\tilde{x}_a^{\text{norm}} \leftarrow \tilde{x}_a \,/\, (\max_t |\tilde{x}_a[t]| + 10^{-8})$
// Scale to 16-bit range
$w \leftarrow \tilde{x}_a^{\text{norm}} \cdot 2^{15}$
// Mel-filterbank extraction
$\mathbf{F} \leftarrow \textsc{MelFilterbank}(w, \text{bins} = 128, sr = 16\text{kHz})$
return $\mathbf{F}$, $\tilde{x}_a^{\text{norm}}$

The mel-filterbank computation uses the Kaldi library with 128 mel bins, 25ms frame length (400 samples), and 10ms frame shift (160 samples) at 16kHz sampling rate. The entire pipeline is differentiable, enabling gradient flow from the loss function to the raw perturbation.

D.2 Gradient Update Step

Algorithm 2 describes a single gradient update step for the perturbation. This modular formulation separates the gradient computation and projection from the outer optimization loop.

The gradient scaling factor γ stabilizes optimization across different loss functions. We use $\gamma \in \{1, 10^{-5}, 10^4\}$ depending on the magnitude of raw gradients.

D.3 Main Optimization Loop

Algorithm 3 describes the outer optimization loop with early stopping. The algorithm iterates over the training dataset, accumulating gradients and tracking the best perturbation found during training.

Training terminates when the loss does not improve for P consecutive epochs. We use $P = 10$ in most experiments. This criterion prevents overfitting to the training set and reduces computational cost for attacks that converge quickly. The best perturbation is saved to disk after each improvement for recovery in case of interruption.

Algorithm 2 Single-Step Gradient Update

Require: Current perturbation δ, sample $(x_a, x_v, q, a^\star)$
Require: Model M_θ, loss function $\mathcal{L}$, budget ε
Require: Gradient scaling factor γ, optimizer state
Ensure: Updated perturbation δ, sample loss ℓ

$\mathbf{F}, \tilde{x}_a \leftarrow \textsc{Preprocess}(x_a, \delta)$ {Algorithm 1}
// Forward pass through frozen model
outputs $\leftarrow M_\theta(\mathbf{F}, x_v, q)$
// Compute adversarial loss
$\ell \leftarrow \mathcal{L}(\text{outputs}, a^\star, x_a, \tilde{x}_a)$
// Backward pass
$\mathbf{g} \leftarrow \nabla_\delta \ell$
// Handle numerical instabilities
$\mathbf{g}[\text{isnan}(\mathbf{g})] \leftarrow 0$
// Apply gradient scaling
$\mathbf{g} \leftarrow \mathbf{g} \cdot \gamma$
// Adam update
$\delta \leftarrow \textsc{AdamUpdate}(\delta, \mathbf{g}, \text{optimizer state})$
// Project onto constraint set
$\delta \leftarrow \text{clip}(\delta, -\varepsilon, \varepsilon)$
return δ, ℓ

Algorithm 3 Main Attack Optimization Loop

Require: Dataset $\mathcal{D} = \{(x_a^{(i)}, x_v^{(i)}, q^{(i)}, a^{\star(i)})\}_{i=1}^{N}$
Require: Model M_θ, loss $\mathcal{L}$, budget ε, learning rate η
Require: Max epochs K, patience P
Ensure: Optimized perturbation δ^{best}

// Initialize perturbation from natural audio
$\delta \leftarrow \text{clip}(\text{BellSound}, -\varepsilon, \varepsilon)$
// Initialize Adam optimizer
optimizer $\leftarrow$ ADAM$(\eta, \beta_1 = 0.9, \beta_2 = 0.999)$
best_loss $\leftarrow \infty$
epochs_no_improve $\leftarrow 0$
for epoch $= 1$ to K **do**
  epoch_loss $\leftarrow 0$
  // Iterate over training samples
  for $(x_a, x_v, q, a^\star) \in \mathcal{D}$ **do**
    $\delta, \ell \leftarrow$ GRADIENTSTEP$(\delta, (x_a, x_v, q, a^\star))$
    epoch_loss $\leftarrow$ epoch_loss $+\ell$
  end for
  epoch_loss $\leftarrow$ epoch_loss / $|\mathcal{D}|$
  // Track best perturbation
  if epoch_loss $<$ best_loss **then**
    best_loss $\leftarrow$ epoch_loss
    $\delta^{\text{best}} \leftarrow \delta$
    epochs_no_improve $\leftarrow 0$
  else
    epochs_no_improve $\leftarrow$ epochs_no_improve $+1$
  end if
  // Early stopping check
  if epochs_no_improve $\geq P$ **then**
    break
  end if
end for
return δ^{best}

E Implementation Details

E.1 Optimization Hyperparameters

All attacks use projected gradient descent with the Adam optimizer. Table 11 summarizes the hyperparameter settings used across experiments.

Hyperparameter	Value
Learning rate η	10^{-4}
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	10^{-8}
Batch size	1
Maximum epochs	150
Early stopping patience	10 epochs
Gradient clipping norm	1.0
Attack budget ε	$\{0.3, 0.5, 0.7, 1.0\}$
Perturbation length	10 seconds
Sampling rate	16 kHz

Table 11: Hyperparameter settings for adversarial attack optimization.

E.2 Perturbation Initialization

The choice of initialization for the perturbation $\delta^{(0)}$ affects both convergence speed and final attack quality. We experimented with two initialization strategies:

(1) Random initialization: Each element of $\delta^{(0)}$ is sampled uniformly from $[-\varepsilon, \varepsilon]$. This provides a generic starting point but may require more iterations to converge.

(2) Natural audio initialization: We initialize $\delta^{(0)}$ using the audio track from a natural video in the dataset. Specifically, we use a recording of bells ringing from the AVQA dataset, clipped to the attack budget ε. This sample is removed from the training set to prevent data leakage. We found that natural audio initialization leads to perturbations that sound more natural when played independently.

Unless specified otherwise, we use natural audio initialization.

E.3 Convergence Conditions

For the attack budget analysis experiments, we allow the optimization to run until convergence rather than fixing a maximum number of epochs. Convergence is defined as no improvement in the training loss for 10 consecutive epochs. Under this criterion, different attack budgets require different numbers of epochs to converge, providing insight into the optimization landscape at different perturbation magnitudes.

E.4 Learning Rate Scheduling

We implement two learning rate scheduling strategies depending on the experiment:

(1) Reduce-on-plateau: The learning rate is reduced by a factor of 0.1 when the loss does not improve for a specified patience window. This is used for experiments where we train until convergence.

(2) Cosine annealing with warmup: For fixed-epoch experiments, we use a warmup phase followed by cosine annealing. The learning rate increases linearly during warmup, then follows a cosine decay schedule:

$$\eta_t = \begin{cases} \eta \cdot \frac{t}{t_{\text{warm}}} & \text{if } t < t_{\text{warm}} \\ \frac{\eta}{2}\left(1 + \cos\left(\pi \cdot \frac{t - t_{\text{warm}}}{T - t_{\text{warm}}}\right)\right) & \text{otherwise} \end{cases} \tag{36}$$

where t_{warm} is the warmup duration and T is the total number of training steps.

Unless specified otherwise, we use reduce-on-plateau.

E.5 Hardware and Computational Resources

All experiments were conducted on NVIDIA GPUs including H200, A100, L40s, and A40s models. The choice of GPU depended on availability and memory requirements of specific experiments. VideoLLAMA2 experiments used full precision (FP32) to ensure numerical stability during gradient computation through the audio preprocessing pipeline. Qwen model experiments used mixed precision (FP16) to reduce memory consumption and accelerate training.

The total computational budget for all experiments reported in this paper is approximately 2,500 GPU-hours. Individual attack training runs range from 20-25 GPU-hours depending on the dataset size, number of epochs, and model complexity. Evaluation on held-out test sets requires additional compute for inference but is substantially faster than training.

E.6 Reproducibility

All experiments use fixed random seeds for reproducibility. We set:

(1) NumPy random seed: 42

(2) PyTorch CUDA seed: 30

These settings ensure that perturbation initialization, data shuffling, and GPU operations produce consistent results across runs. We found that attack success rates are sensitive to random initialization, making seed fixing essential for reproducible comparisons.

F Evaluation Metrics

This section provides detailed definitions and interpretations for all evaluation metrics used in our experiments.

F.1 Attack Success Rate

The Attack Success Rate (ASR) quantifies the fraction of originally correct predictions that are flipped to incorrect predictions after applying the adversarial perturbation. This metric specifically measures the attack's ability to induce failures on samples where the model would otherwise succeed, excluding samples where the model was already incorrect on clean inputs.

Formally, given an evaluation dataset $\mathcal{D} = \{(x_a^{(i)}, x_v^{(i)}, q^{(i)}, a^{\star(i)})\}_{i=1}^{N}$, we define:

$$\text{CleanCorrect}^{(i)} = \mathbf{1}\left[\hat{a}(x_a^{(i)}, x_v^{(i)}, q^{(i)}; \theta) = a^{\star(i)}\right], \tag{37}$$

where $\hat{a}(\cdot)$ denotes the model's predicted answer and $\mathbf{1}[\cdot]$ is the indicator function. Similarly, for perturbed audio $\tilde{x}_a^{(i)} = x_a^{(i)} + \delta$:

$$\text{AttackSuccess}^{(i)} = \mathbf{1}\Big[\text{CleanCorrect}^{(i)} = 1 \wedge \hat{a}(\tilde{x}_a^{(i)}, x_v^{(i)}, q^{(i)}; \theta) \neq a^{\star(i)}\Big]. \tag{38}$$

The ASR is then computed as:

$$\text{ASR} = \frac{\sum_{i=1}^{N} \text{AttackSuccess}^{(i)}}{\sum_{i=1}^{N} \text{CleanCorrect}^{(i)}}. \tag{39}$$

This definition ensures that ASR measures genuine attack-induced failures rather than being inflated by baseline model errors. An ASR of 100% indicates that the attack successfully flips all correct predictions to incorrect ones.

F.2 LPIPS for Audio

We adapt the Learned Perceptual Image Patch Similarity (LPIPS) metric to the audio domain by computing it on log-mel spectrogram representations. Given clean audio x_a and perturbed audio $\tilde{x}_a$, we first compute their log-mel spectrograms $\mathbf{S}$ and $\tilde{\mathbf{S}}$ respectively. These spectrograms are treated as single-channel images and fed to a pretrained neural network such as AlexNet (Krizhevsky et al., 2012) or VGG (Simonyan and Zisserman, 2015) to extract deep features. LPIPS computes the distance between activations at multiple layers of the network:

$$\text{LPIPS}(x_a, \tilde{x}_a) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|\Delta_{l,hw}\|_2^2, \quad \Delta_{l,hw} = w_l \odot \Big(\phi_l(\mathbf{S})_{hw} - \phi_l(\tilde{\mathbf{S}})_{hw}\Big), \tag{40}$$

where ϕ_l extracts features at layer l, w_l are learned weights, and the sum is over spatial positions (h, w) in the feature maps of height H_l and width W_l.

Unlike pixel-wise metrics such as MSE, LPIPS captures perceptual differences by comparing high-level features. Lower LPIPS values indicate higher perceptual similarity between clean and perturbed audio, while higher values correspond to increasingly noticeable distortions. This comparison is motivated by Chary and Ramirez (2025) where authors use LPIPS for audio reconstruction.

F.3 Scale-Invariant Signal-to-Noise Ratio

Scale-Invariant Signal-to-Noise Ratio (SI-SNR) measures the similarity between a reference signal and an estimated signal while removing any global gain differences. This property makes SI-SNR more robust than standard SNR for evaluating audio perturbations, as it focuses on waveform shape rather than overall amplitude.

Given clean signal $\mathbf{x} \in \mathbb{R}^T$ and perturbed signal $\hat{\mathbf{x}} \in \mathbb{R}^T$, SI-SNR is computed as follows. First, both signals are zero-meaned:

$$\mathbf{x} \leftarrow \mathbf{x} - \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} - \frac{1}{T}\sum_{t=1}^{T} \hat{x}_t. \tag{41}$$

Next, we project $\hat{\mathbf{x}}$ onto $\mathbf{x}$ to obtain the signal-aligned component:

$$\mathbf{x}_{\text{target}} = \frac{\langle \hat{\mathbf{x}}, \mathbf{x} \rangle}{\|\mathbf{x}\|^2} \mathbf{x}. \tag{42}$$

The residual error orthogonal to $\mathbf{x}$ is:

$$\mathbf{e}_{\text{noise}} = \hat{\mathbf{x}} - \mathbf{x}_{\text{target}}. \tag{43}$$

Finally, SI-SNR is the logarithmic ratio of signal energy to noise energy:

$$\text{SI-SNR}(\hat{\mathbf{x}}, \mathbf{x}) = 10 \log_{10} \left(\frac{\|\mathbf{x}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{noise}}\|^2} \right). \tag{44}$$

Higher SI-SNR values indicate closer alignment between perturbed and clean signals. A value of 0 dB means the noise energy equals the signal energy. Negative values indicate that the perturbation dominates the original signal, though the original audio may still be audible if the perturbation is structured rather than random.

F.4 Word Error Rate

For evaluating attacks on automatic speech recognition systems, we use Word Error Rate (WER), the standard metric for transcription accuracy:

$$\text{WER} = \frac{S + D + I}{N} \times 100\%, \tag{45}$$

where S is the number of word substitutions, D is the number of deletions, I is the number of insertions required to transform the predicted transcript into the reference, and N is the total number of words in the reference transcript.

We report both absolute WER on attacked audio and the change in WER (ΔWER = WER(attacked) $-$ WER(clean)) to quantify attack impact relative to baseline performance.

G LLM-as-a-Judge Evaluation

For evaluating model predictions on the AVSD dataset, we employ an LLM-as-a-judge approach that automates the assessment of free-form text generation quality. This section describes the evaluation protocol and provides the complete prompt template.

G.1 Evaluation Protocol

The evaluation proceeds through the following steps:

(1) For each sample in the AVSD validation set, we collect the ground truth caption, summary, and all conversation Q&A turns that describe the video content.

(2) We generate predictions using both clean audio and adversarially perturbed audio, obtaining two descriptions for each video.

(3) Each prediction is independently evaluated by the LLM judge against the ground truth information. The judge assesses factual accuracy, consistency with visual content, and absence of hallucinations.

(4) The judge outputs a structured response containing: a brief analysis explaining the reasoning, a binary verdict (CORRECT or INCORRECT), and a confidence level (HIGH, MEDIUM, or LOW).

(5) We compute the attack success rate as the fraction of samples where the clean prediction was judged CORRECT but the adversarial prediction was judged INCORRECT.

This automated evaluation enables consistent and scalable assessment across the full validation set while maintaining alignment with human judgment criteria. The structured output format allows for reliable parsing of verdicts without manual inspection.

G.2 Prompt Template

Figure 5 presents the complete prompt template used for LLM-as-a-judge evaluation.

G.3 Judge Model Selection

We use the `openai/gpt-oss-20b` model as the evaluation judge. This model was selected for its strong instruction-following capabilities and consistent output formatting. The temperature is set to 0 to ensure deterministic evaluations across samples.

LLM Judge Prompt Template

You are an expert evaluator for video description quality assessment. Your task is to determine whether a model's prediction accurately describes a video based on ground truth information.

=== GROUND TRUTH INFORMATION ===
Video Caption: {gt_caption}
Video Summary: {gt_summary}
Detailed Q&A about the video:
{conversation_context}

=== MODEL PREDICTION TO EVALUATE ===
{prediction}

=== EVALUATION CRITERIA ===
Evaluate the prediction on these dimensions:

1. **Subject Identification**: Does the prediction correctly identify the main person/people and their gender/age if mentioned?
2. **Action Accuracy**: Does the prediction correctly describe what the person is doing (actions, activities)?
3. **Object Recognition**: Are the objects mentioned in the prediction consistent with ground truth (e.g., food items, furniture, clothing)?
4. **Setting/Location**: Is the location/setting correctly identified (room type, indoor/outdoor)?
5. **Factual Consistency**: Are there any direct contradictions or hallucinations compared to ground truth?

=== JUDGMENT GUIDELINES ===
Mark as **CORRECT** if:

- The prediction captures the essential scene accurately
- Main actions and subjects are correctly identified
- No major factual errors or hallucinations
- Minor omissions or paraphrasing are acceptable

Mark as **INCORRECT** if:

- The prediction misidentifies the main subject or their actions
- There are factual contradictions with ground truth
- The prediction describes a substantially different scene
- Critical details are wrong (e.g., wrong objects, wrong location type)

=== OUTPUT FORMAT ===
Provide your response in this EXACT format:
ANALYSIS: [Briefly explain your reasoning in 2-3 sentences, comparing key elements]
VERDICT: [CORRECT/INCORRECT]
CONFIDENCE: [HIGH/MEDIUM/LOW]

Figure 5: Prompt template used for LLM-as-a-judge evaluation on AVSD. The judge evaluates predictions against ground truth captions, summaries, and conversation context.

H Additional Results and Numerical Tables

This section provides exact numerical values for all figures and additional analysis presented in the main paper.

H.1 Training Iterations and Attack Success Rate

Table 12 reports the relationship between total training iterations, dataset size, and attack success rate for $\mathcal{L}_{\text{negLM}}$. These results correspond to Figure 2 in the main paper.

Total Iterations	ASR (%)	Dataset Size
200k	1.93	20k
400k	73.64	20k
660k	75.18	20k
100k	2.16	10k
200k	78.82	10k
500k	93.15	10k
100k	5.91	2k
200k	6.84	2k
300k	10.27	2k
406k	72.3	2k
121k	36.76	1k
200k	64.07	1k
372k	87.00	1k
390k	78.4	1k

Table 12: Training iterations, dataset size, and attack success rate for $\mathcal{L}_{\text{negLM}}$ on VideoLLAMA2. Extended optimization on smaller datasets yields higher ASR than brief training on larger datasets.

The results reveal that attack effectiveness depends more on total optimization iterations than on dataset diversity. Configurations with only 1,000 training samples but 372,000 iterations achieve 87% ASR, substantially outperforming configurations with 20,000 samples but only 200,000 iterations (1.93% ASR). This pattern suggests that adversarial perturbations exploit consistent model behaviors that can be identified through repeated optimization on a small, representative subset of the data. Training with 20k data samples converges on 660k iterations, training with 10k data samples converges on 500k iterations, training with 2k data samples converges on 406k iterations, and training with 1k data samples converges on 390k iterations.

H.2 Perceptibility Analysis

Table 13 reports perceptual distortion metrics alongside attack effectiveness for each objective. These values correspond to Figure 4.

Attack Objective	ΔWER	LPIPS	SI-SNR (dB)
$\mathcal{L}_{\text{negLM}}$	0.972	0.22	−11.48
$\mathcal{L}^{(\text{cos})}$	0.959	0.08	−1.77
$\mathcal{L}^{(\text{visionatt})}$	0.100	0.07	−1.02
$\mathcal{L}^{(\text{audioatt})}$	0.047	0.06	0.33
$\mathcal{L}^{(\text{randatt})}$	0.062	0.06	0.05
$\mathcal{L}^{(\text{hidden-cos})}$	0.036	0.06	0.56
$\mathcal{L}^{(\text{combined})}$	0.491	0.14	−6.23

Table 13: Perceptual distortion metrics and WER difference for each attack objective. Higher ΔWER correlates with higher LPIPS, indicating that speech recognition models respond primarily to acoustic distortion magnitude.

The encoder-space attack $\mathcal{L}^{(\text{cos})}$ achieves a favorable trade-off between attack effectiveness and perceptibility. Despite inducing nearly as much damage to Whisper as $\mathcal{L}_{\text{negLM}}$ (0.959 vs 0.972 ΔWER), it maintains substantially lower perceptual distortion (LPIPS 0.08 vs 0.22, SI-SNR -1.77 vs -11.48 dB). This indicates that the encoder-space attack exploits structured vulnerabilities in the audio representation rather than relying on brute-force signal corruption.

H.3 Attention Allocation Analysis

Table 14 reports the average attention mass assigned to audio and video tokens under different attack conditions. Values are normalized by the number of layers and tokens to enable comparison across models.

Objective	VL2-audio	VL2-video	Q2.5-audio	Q2.5-video
Clean	9.01	**15.50**	13.01	**39.04**
$\mathcal{L}_{\text{negLM}}$	9.55	14.94	12.96	31.86
$\mathcal{L}^{(\text{cos})}$	13.55	13.96	13.03	35.00
$\mathcal{L}^{(\text{visionatt})}$	14.03	13.49	**13.50**	34.42
$\mathcal{L}^{(\text{audioatt})}$	21.49	13.85	12.88	32.73
$\mathcal{L}^{(\text{randatt})}$	10.51	14.65	13.03	32.48
$\mathcal{L}^{(\text{hidden-cos})}$	13.55	14.43	13.04	34.51
$\mathcal{L}^{(\text{combined})}$	**29.23**	13.30	13.17	34.34

Table 14: Average attention logits assigned to audio and video tokens for VideoLLAMA2 (VL2) and Qwen 2.5 Omni (Q2.5), normalized by layer and token count. The combined attack produces the largest shift in audio-token attention on VideoLLAMA2 (29.23 vs 9.01 clean).

H.4 Sequence Confidence Under Attack

Table 15 reports the average sequence confidence of generated outputs under clean and adversarial conditions. Confidence is computed as the exponential of the mean log-probability of all generated tokens.

The small confidence differences under attack are notable. For example, the combined attack on AVQA achieves 96.03% ASR while the confidence difference is only 0.080 (0.930 clean vs 0.850 adversarial). This indicates that the model produces incorrect answers with nearly the same confidence as correct answers, making adversarial failures difficult to detect through output uncertainty monitoring.

H.5 Clean versus Attacked Accuracy

Table 16 reports model accuracy on clean and adversarial audio across AVQA and Music-AVQA for all attack objectives.

The results confirm that high ASR values reflect genuine attack-induced failures rather than poor baseline performance. The model achieves 95.6% accuracy on clean AVQA inputs. Under the combined attack, accuracy drops to 3.9%, representing a 91.8 percentage point degradation that is attributable entirely to the adversarial perturbation.

I Layer-wise Attention Analysis

This section provides detailed layer-wise analysis of attention patterns under different attack conditions. Understanding how attacks affect attention at different depths of the network provides insight into the mechanisms by which adversarial perturbations induce failures.

I.1 Summary Statistics

Tables 17 and 18 report summary statistics for attention allocation to audio and video tokens across the 28 layers of VideoLLAMA2.

I.2 Layer-wise Attention Patterns

Figure 6 visualizes the average attention mass assigned to audio and video tokens at each transformer layer under different attack objectives. Several patterns emerge from this analysis.

Video attention exhibits a characteristic front-loaded pattern across all conditions: attention is highest in early layers (1–5), then stabilizes at lower values through middle and late layers. This pattern reflects the model's processing of visual information, where early layers encode low-level features and later layers integrate them with other modalities.

Audio attention behaves differently depending on the attack objective. Under clean conditions and weak attacks ($\mathcal{L}_{\text{negLM}}$, $\mathcal{L}^{(\text{randatt})}$), audio attention remains relatively flat across layers. However, attacks that explicitly target audio pathways ($\mathcal{L}^{(\text{audioatt})}$, $\mathcal{L}^{(\text{combined})}$) induce a global upward shift in audio-token attention that is uniform across all layers. This uniform elevation suggests that these attacks modify how the model processes audio information at a fundamental level, rather than targeting specific computational stages.

The combined attack produces the most dramatic effect, with audio attention elevated to 25–30 across all layers compared to 5–17 under clean conditions. This three-fold increase in audio attention correlates with the attack's strong performance (96.03% ASR), supporting the hypothesis that forcing the model to attend heavily to corrupted audio is an effective attack strategy.

J Perturbation Visualization

Figure 7 visualizes the waveforms of adversarial audio obtained after optimizing different loss functions, overlaid on the clean reference signal. Understanding the temporal structure of learned perturbations provides insight into the attack mechanisms.

J.1 Initialization and Structure

All perturbations are initialized using a natural bell-ringing sound sampled from the AVQA dataset. The corresponding video is explicitly excluded from training to prevent data leakage. This initialization provides a structured starting point with temporal patterns that can be refined through optimization. The bell sound contains a mixture of transient attacks (the initial strike) and sustained resonances (the ringing decay), providing diverse spectral content.

J.2 Perturbation Characteristics

Across loss functions, the learned perturbations exhibit several common characteristics:

(1) Low amplitude: Perturbations remain visually similar to the clean signal in waveform plots, with amplitudes constrained by the attack budget ε.

(2) Temporal smoothness: Unlike random noise, the learned perturbations maintain temporal coherence inherited from the natural audio initialization.

Objective	AVQA-Adv	AVQA-Clean	Δ	M-AVQA-Clean	M-AVQA-Adv	Δ	AVSD-Adv	AVSD-Clean	Δ
$\mathcal{L}_{\text{negLM}}$	0.756	0.816	0.060	0.752	0.818	0.066	0.640	0.650	0.010
$\mathcal{L}^{(\text{cos})}$	0.800	0.930	0.130	0.741	0.823	0.082	0.650	0.660	0.010
$\mathcal{L}^{(\text{visionatt})}$	0.838	0.935	0.097	0.765	0.798	0.033	0.660	0.650	−0.010
$\mathcal{L}^{(\text{audioatt})}$	0.712	0.935	0.223	0.770	0.813	0.043	0.660	0.650	−0.010
$\mathcal{L}^{(\text{randatt})}$	0.713	0.798	0.085	0.770	0.816	0.046	0.650	0.650	0.000
$\mathcal{L}^{(\text{hidden-cos})}$	0.822	0.930	0.108	0.737	0.800	0.063	0.650	0.640	−0.010
$\mathcal{L}^{(\text{combined})}$	0.850	0.930	0.080	0.750	0.855	0.105	0.642	0.650	0.008

Table 15: Sequence confidence for clean and adversarial settings across datasets, computed for successful attacks only. Δ = Clean − Adv. Small or negative differences indicate that adversarial failures arise from internal misalignment rather than model uncertainty.

Objective	AVQA Clean	AVQA Attack	Δ	Music-AVQA Clean	Music-AVQA Attack	Δ
$\mathcal{L}_{\text{negLM}}$	0.956	0.850	0.106	0.807	0.743	0.064
$\mathcal{L}^{(\text{cos})}$	0.956	0.110	0.846	0.807	0.747	0.060
$\mathcal{L}^{(\text{visionatt})}$	0.956	0.920	0.036	0.807	0.777	0.030
$\mathcal{L}^{(\text{audioatt})}$	0.956	0.410	0.546	0.807	0.769	0.038
$\mathcal{L}^{(\text{randatt})}$	0.956	0.910	0.046	0.807	0.758	0.049
$\mathcal{L}^{(\text{hidden-cos})}$	0.956	0.795	0.161	0.807	0.766	0.041
$\mathcal{L}^{(\text{combined})}$	0.956	0.039	**0.918**	0.807	0.750	**0.057**

Table 16: Accuracy comparison between clean and adversarial audio. Δ denotes the accuracy drop (Clean − Attack). High ASR is driven by substantial attack-induced degradation, not low baseline accuracy.

Attack Objective	Mean	Max	Min
$\mathcal{L}_{\text{negLM}}$	9.55	17.77	4.83
$\mathcal{L}^{(\text{cos})}$	13.55	26.36	8.08
$\mathcal{L}^{(\text{visionatt})}$	14.03	26.47	8.57
$\mathcal{L}^{(\text{audioatt})}$	21.49	26.24	19.50
$\mathcal{L}^{(\text{randatt})}$	10.51	20.91	3.36
$\mathcal{L}^{(\text{hidden-cos})}$	13.55	25.06	5.42
$\mathcal{L}^{(\text{combined})}$	**29.23**	30.34	25.96

Table 17: Summary statistics for audio token attention across 28 layers of VideoLLAMA2. The combined attack produces uniformly elevated audio attention (mean 29.23) with small variance across layers.

Attack Objective	Mean	Max	Min
$\mathcal{L}_{\text{negLM}}$	14.94	34.64	6.12
$\mathcal{L}^{(\text{cos})}$	13.96	30.73	5.82
$\mathcal{L}^{(\text{visionatt})}$	13.49	29.88	5.30
$\mathcal{L}^{(\text{audioatt})}$	13.85	29.86	5.49
$\mathcal{L}^{(\text{randatt})}$	14.65	34.81	5.96
$\mathcal{L}^{(\text{hidden-cos})}$	14.43	29.66	6.53
$\mathcal{L}^{(\text{combined})}$	13.30	28.34	5.23

Table 18: Summary statistics for video token attention across 28 layers of VideoLLAMA2. Video attention remains relatively stable across attack conditions compared to audio attention.

(3) Structured deviation: Effective attacks ($\mathcal{L}^{(\text{cos})}$, $\mathcal{L}^{(\text{combined})}$) introduce small but consistent modifications that accumulate across time, rather than isolated high-amplitude spikes.

The attention-based attacks ($\mathcal{L}^{(\text{audioatt})}$, $\mathcal{L}^{(\text{hidden-cos})}$) produce the least visible deviation from the clean signal, yet still achieve measurable attack success. This reinforces the finding that structured, low-distortion perturbations can be more effective than high-distortion approaches when targeting specific computational mechanisms.

K Realism of Threat Model and Societal Implications

We discuss the practical realism of our threat model and its broader implications. Our attack operates under an untargeted, audio-only setting, where a shared perturbation is learned across samples. This design avoids assumptions required by targeted or per-sample attacks, such as prior knowledge of inputs or predefined malicious outputs. Instead, the attack induces internal representation shifts that lead to incorrect predictions, which are harder to detect in generative settings.

A key aspect of realism lies in the single-

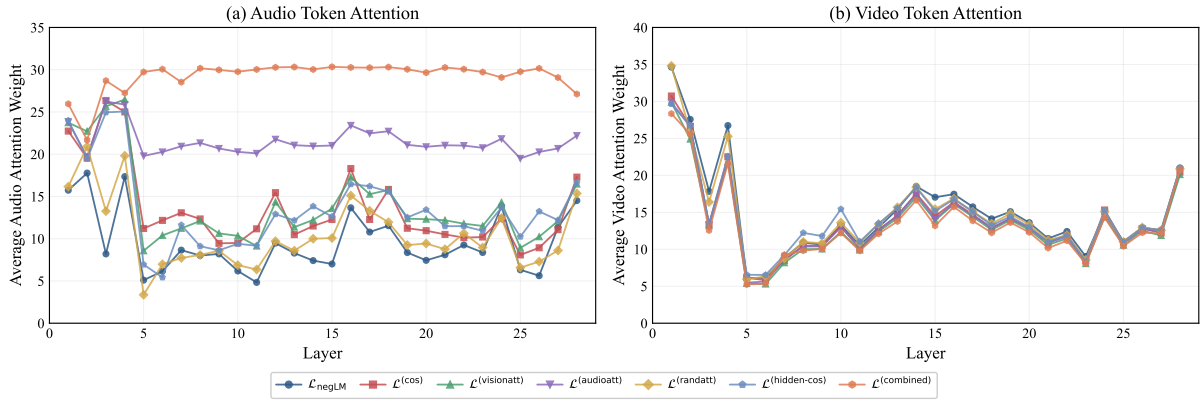

Figure 6: Layer-wise attention analysis for VideoLLAMA2. Left: Average attention mass on audio tokens per layer. Right: Average attention mass on video tokens per layer. Different attack objectives induce distinct attention patterns across the 28 transformer layers. The combined attack ($\mathcal{L}^{(\text{combined})}$) shows uniformly elevated audio attention across all layers.

modality constraint. Unlike prior multimodal attacks that require simultaneous manipulation of multiple inputs, our method assumes access only to the audio channel, which is significantly easier to control in real-world environments. Audio perturbations can be delivered through speakers, embedded in media, or introduced as background noise, making them difficult to detect and filter.

We further extend our evaluation to physically motivated settings by incorporating room impulse response (RIR) transformations during training. This allows the perturbation to account for environmental effects such as reverberation and signal degradation. As shown in 7, RIR-based training improves robustness of the attack under such transformations, although physical attacks generally remain weaker due to environmental variability.

Societal Implications. Audio-only vulnerabilities introduce risks in applications where multimodal reasoning is critical. These include assistive systems, surveillance pipelines, autonomous agents, and human-AI interaction platforms. Adversarial audio signals could be embedded in public environments or media streams, leading to incorrect system behavior without obvious perceptual cues. This highlights the need for defenses that enforce cross-modal consistency and improve robustness to adversarial perturbations in realistic acoustic conditions.

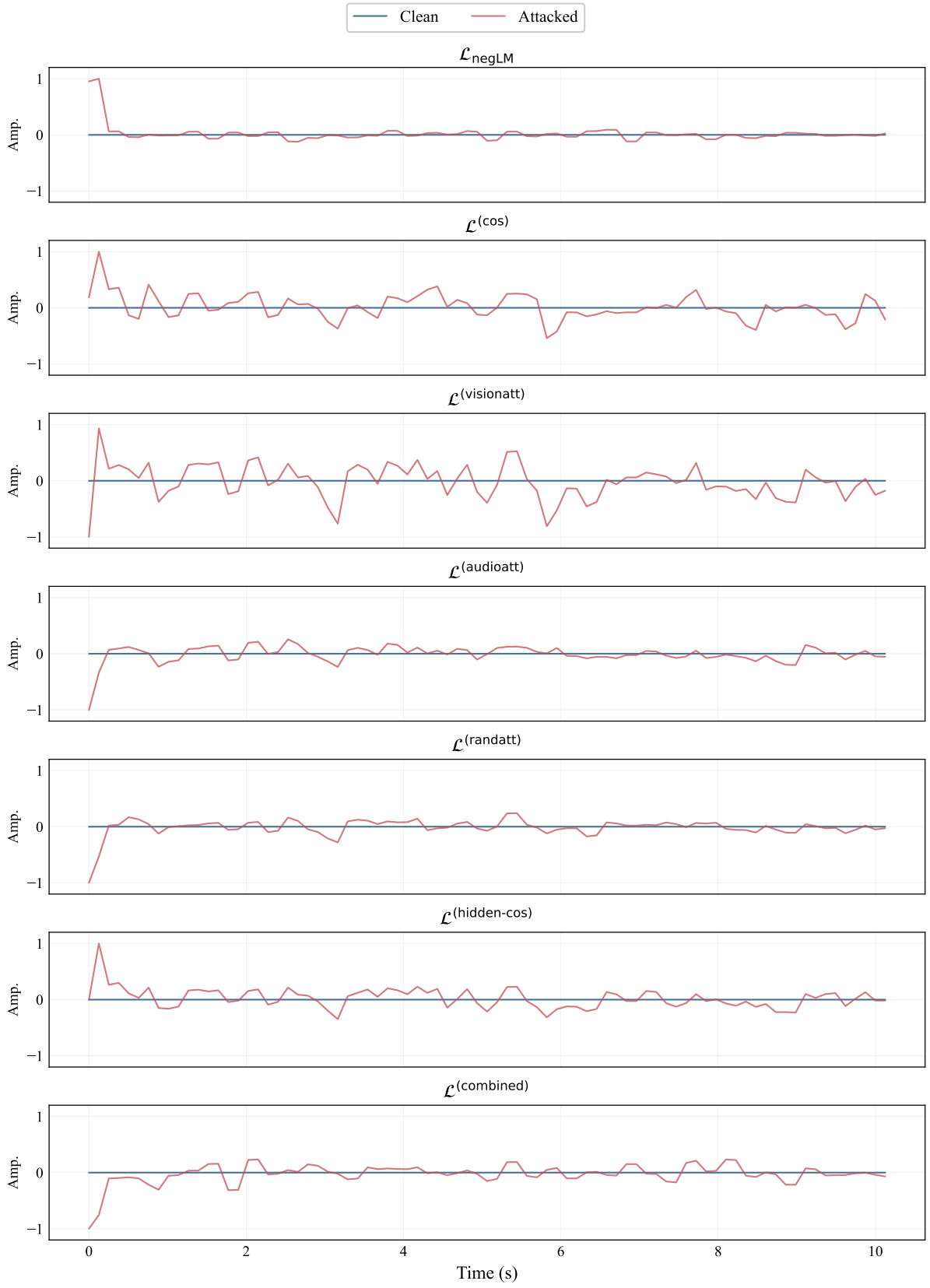

Figure 7: Waveform visualization of learned audio-only adversarial perturbations. Clean (blue) and attacked (red) audio waveforms are overlaid for perturbations optimized using different loss functions. Sampling rate is reduced for visualization clarity. Effective attacks produce perturbations that remain structurally similar to the clean signal.