SOUNDBREAK: Exploring the Effectiveness of Audio Attacks on Audio-Video Models

Anonymous ACL submission

Abstract

Recent multimodal foundation models that jointly process audio, video, and text have shown remarkable capabilities but remain vulnerable to adversarial attacks. We present SOUNDBREAK, a comprehensive study of audio-based adversarial attacks on trimodal audio-video models that reveals critical vulnerabilities through the audio channel alone. Unlike existing works that target single or dual modalities, we show that **carefully crafted audio perturbations can disrupt model behavior**, achieving attack success rates up to **89%**. We systematically evaluate **six** distinct attack strategies ranging from negative language modeling to attention manipulation across **three** models on the AVQA and AVSD benchmarks. Our results reveal that encoder-space attacks targeting audio embeddings prove 6× more effective than layer-wise perturbations. Critically, we find that attacks exhibit **limited transferability** across models due to architectural differences-with reasoning-focused models showing 67× lesser magnitude updates for same training durations. Our work reveals fundamental vulnerabilities in current multimodal architectures and their limitations in ensuring robustness under real-world deployment scenarios.

1 Introduction

Recent progress in multimodal large language models has intensified research on their susceptibility to attacks across multiple modalities. (Liu et al., 2025b). While these models demonstrate remarkable capabilities in understanding complex multimodal interactions, several works have exposed vulnerabilities in CLIP-based and vision-language models (Zhang et al., 2025b; Mei et al., 2025; Xu et al., 2024), with growing interest in audio-based attacks (Raina et al., 2024). Despite this progress, robustness against adversarial manipulation remains largely unexplored, particularly in scenarios where attackers can only control a single modality among more than two input modalities.

Attacks on multimodal models generally fall into four categories: *prompt injection, jailbreaks, data poisoning, and adversarial perturbations* (Liu et al., 2025b). Jailbreaks bypass safety constraints through prompt modification (Yi et al., 2024) or gradient-based optimization (Wang et al., 2024), while prompt injection embeds harmful instructions directly into inputs (Bagdasaryan et al., 2023). Data poisoning and backdoor attacks compromise training pipelines (Walmer et al., 2022). Adversarial perturbations, however, present unique cross-modal risks affecting interpretations without directly altering those modalities (Kang et al., 2025a). These vulnerabilities underscore a fundamental threat that existing research has yet to address.

The current landscape of adversarial attacks on multimodal models reveals significant gaps in our understanding. Previous work has predominantly focused on attacking vision-language models through image perturbations (Zhang et al., 2025b; Mei et al., 2025; Hao et al., 2025), with researchers demonstrating that carefully crafted visual noise can manipulate model outputs. Similarly, recent studies have explored audio-specific attacks on speech models (Raina et al., 2024; Ma et al., 2025; Sadasivan et al., 2025), showing that acoustic adversarial examples can silence or redirect speech recognition systems. However, these approaches treat modalities in isolation, failing to capture the complex interplay between audio, vision, and language in truly multimodal systems.

The few attempts at multimodal attacks have relied on simultaneous manipulation of multiple input channels (Mustakim et al., 2025) or require access to training data for backdoor insertion (Walmer et al., 2022). These approaches face practical limitations: coordinating attacks across multiple modalities is technically challenging and easily detectable, while training-time attacks assume unrealistic access to model development pipelines. Furthermore, existing targeted attacks that aim for

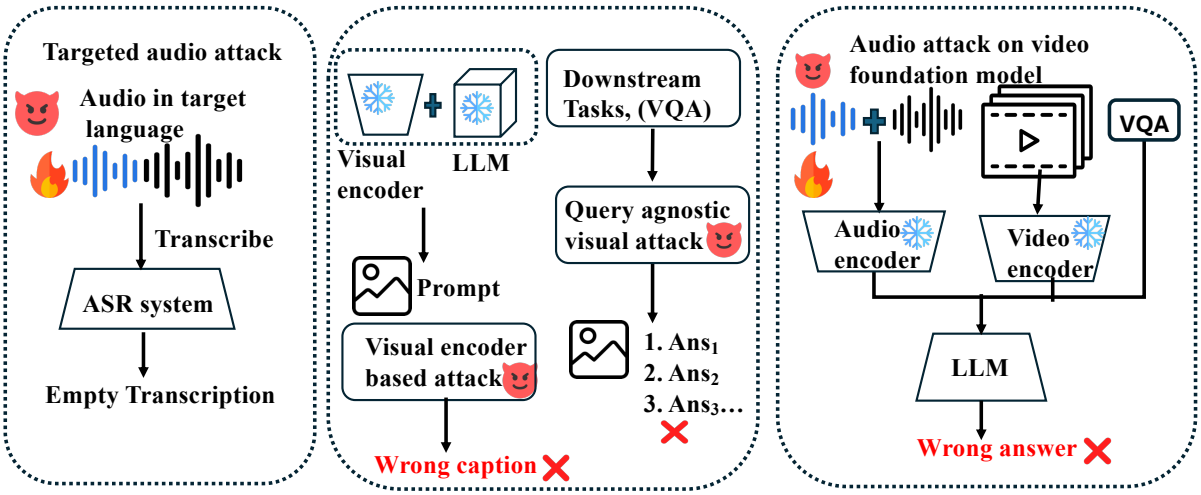

Figure 1: Progression of adversarial attacks from audio and vision-language systems to our untargeted audio attack on a trimodal video–language model, where perturbations in the audio channel alone induce incorrect outputs.

specific malicious outputs create obvious behavioral anomalies that modern monitoring systems can readily identify.

Our work addresses these critical gaps by investigating a more realistic and critical threat model: **untargeted audio-only attacks on audio-video models**. We show that an attacker who controls only the audio channel can systematically degrade model performance across all modalities through gradient-based optimization of adversarial perturbations. This approach is particularly concerning because audio manipulation is often easier to execute in practice-through compromised microphones, environmental speakers, or transmission channels-while being harder to detect than visual artifacts (Choi et al., 2024).

To summarize, we make three contributions: **(1)** We propose six distinct loss functions for audio-based attacks, each targeting different aspects of the model's architecture. **(2)** We conduct extensive experiments across three state-of-the-art audio-visual models using standardized audio-visual tasks, revealing model-specific vulnerabilities and the limited transferability of attacks. **(3)** We analyze the relationship between attack methodologies, transferability, training duration, and data efficiency, providing practical insights for both attackers and defenders.

2 Related Work

Foundations of Adversarial Attacks The discovery of adversarial examples fundamentally challenged our understanding of neural network robustness (Szegedy et al., 2013; Nguyen et al., 2015). Early investigations revealed the limitations of deep learning in adversarial settings (Papernot et al., 2016), while the Fast Gradient Sign Method demonstrated how linear behavior in high-dimensional spaces enables these vulnerabilities (Goodfellow et al., 2015a,b). The development of sophisticated attack methods progressed from DeepFool's geometric approach (Moosavi-Dezfooli et al., 2016) to the powerful optimization-based C&W attack (Carlini and Wagner, 2017), proving that adversarial robustness requires fundamental architectural changes rather than superficial defenses. Recent advances have successfully adapted these techniques to large language models through projected gradient descent (Geisler et al., 2024), achieving order-of-magnitude improvements in computational efficiency while maintaining attack effectiveness.

Vision-Centric Multimodal Attacks The emergence of vision-language models has spawned a new generation of adversarial attacks that exploit the vision-text interface. Query-agnostic attacks show that a single adversarial image can cause failures across multiple unspecified questions by targeting the vision-language alignment module (Zhang et al., 2025b). This vulnerability extends to downstream-agnostic scenarios where attacking only the vision encoder suffices to degrade performance across diverse tasks (Mei et al., 2025). Recent work has shown that scenario-matched adversarial images combined with multi-loss optimization can achieve jailbreak success rates exceeding 80% on commercial models (Hao et al., 2025). The scalability of these attacks has been

addressed through self-supervised pre-training on large-scale datasets, enabling any-to-any targeted manipulations (Zhang et al., 2025a).

Audio-Domain Adversarial Techniques The audio modality presents unique challenges for adversarial attacks. Universal acoustic segments can effectively "mute" speech models by mimicking end-of-text tokens (Raina et al., 2024), while targeted perturbations enable flexible control over model outputs through selective activation based on speaker characteristics (Ma et al., 2025). Physical-world robustness has been demonstrated through over-the-air attacks that remain effective despite environmental distortions (Sadasivan et al., 2025). Recent work on speech translation systems reveals that both imperceptible perturbations and adversarially crafted music can inject malicious translations (Liu et al., 2025a). Defense mechanisms like SPIRIT show promise through activation patching but highlight the fundamental tension between robustness and utility (Djanibekov et al., 2025). The stealthy jailbreak framework AdvWave demonstrates that gradient shattering can be overcome through dual-phase optimization while maintaining perceptual naturalness (Kang et al., 2025b).

Our work extends beyond these unimodal and bimodal approaches by systematically investigating untargeted attacks that affect only the audio input modality in trimodal settings, revealing how single-channel manipulation can compromise entire multimodal systems through cross-modal vulnerabilities that existing defenses fail to address.

3 Problem Statement

Given an audio-video model $M_\theta : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{Q} \to \mathcal{P}(\mathcal{A})$ 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, we consider 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 norm constraints with ε being the attack budget. Under the white-box assumption the attacker has full access to θ, model outputs and internal differentiable representations, and may therefore evaluate gradients $\nabla_{x_a} \mathcal{J}$ for each chosen loss index $i \in \{1, \dots, k\}$ we define the iterative refinement of the additive perturbation by

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

where $\delta^{(0)} \in \mathcal{C}$, $\Pi_{\mathcal{C}}$ denotes projection onto the feasible set $\mathcal{C}$, $\eta > 0$ is a step size, and $s_i \in \{+1, -1\}$ is a sign chosen so that the same descent-style update implements the intended direction for the i-th loss (i.e., depending on whether the attacker should maximize or minimize $\mathcal{L}^{(i)}$). After K iterations the attack uses $\tilde{x}_a = x_a + \delta^{(K)}$ as the perturbed audio input.

Depending on the nature of $\mathcal{L}^{(i)}$ we choose s_i to indicate whether the optimization should drive the loss up or down; the update above leaves these implementation details implicit and focuses on the general iterative projection-based refinement.

4 Threat model

Attacker's goals

The attacker's objective is to cause the multi-modal model $M_\theta : \mathcal{X}_a \times \mathcal{X}_v \times \mathcal{Q} \to \mathcal{P}(\mathcal{A})$ to produce misleading responses when presented with a legitimate video x_v and a perturbed audio input $\tilde{x}_a$. Concretely, under an *untargeted* threat the attacker seeks to change the model's output distribution so that the produced answer is any response other than the original (honest) one:

$$p(\cdot \mid \tilde{x}_a, x_v, q; \theta) \neq p(\cdot \mid x_a, x_v, q; \theta).$$

Attacker's capabilities

We assume a white-box, audio-only additive adversary with the following capabilities:

- **Full model access.** The attacker has access to model internals (architecture and weights) and may query or compute model outputs and internal differentiable representations.
- **Audio-only perturbation.** The attacker may only modify the audio channel, producing

  $$\tilde{x}_a = x_a + \delta,$$

  where $\delta \in \mathbb{R}^T$ is the additive perturbation. The video input x_v and the question q remain unchanged.

- **Constrained perturbations.** Perturbations are constrained to a feasibility set $\mathcal{C}$, e.g. $\mathcal{C} = \{\delta : \|\delta\|_p \leq \varepsilon\}$, and the attacker can project iterates onto $\mathcal{C}$ during optimization.

5 Attack Methodology

We present six attack strategies that target different computational aspects of trimodal models, from high-level language modeling to low-level attention mechanisms learned through projected gradient descent (PGD).

5.1 Negative Language Modeling based Adversarial Loss

A simple but effective choice of objective for the adversary is to directly reduce the probability of the model's correct answer under the perturbed audio. To achieve this, we define a loss function that is the negative of the standard language modeling loss.

Let $a^\star \in \mathcal{A}$ denote the ground-truth answer associated with input (x_a, x_v, q). The standard language modeling loss for this sample is given by the negative log-likelihood

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

To encourage the perturbation δ to *reduce* the model's confidence in the correct answer, we instead take the **negative** of this loss as the adversarial objective:

$$\begin{aligned}&\mathcal{L}^{(i)}_{\text{negLM}}\big(p(\cdot \mid \tilde{x}_a, x_v, q; \theta), a^\star\big)\\ &= \log p(a^\star \mid \tilde{x}_a, x_v, q; \theta)\,.\end{aligned}$$

Within the iterative refinement procedure defined in 3, this corresponds to choosing the loss index i to point to $\mathcal{L}^{(i)}_{\text{negLM}}$ and setting the sign $s_i = -1$. In this way, the update step

$$\delta^{(k+1)} = \Pi_{\mathcal{C}}\Big(\delta^{(k)} - \eta\, s_i\, \nabla_\delta\, \mathcal{L}^{(i)}_{\text{negLM}}\Big) \qquad (3)$$

acts to reduce the probability mass assigned to the correct answer $a^\star$, thereby implementing an untargeted adversarial attack that destabilizes the model's prediction.

5.2 Encoder-based Cosine Similarity Reduction Based Loss

Since the attack primarily focuses on audio, we target the audio encoder to perform our adversarial attack. Let $f_a : \mathcal{X}_a \to \mathbb{R}^d$ denote the audio encoder of M_θ, which maps an audio waveform to a d-dimensional feature embedding. Specifically, we define a loss that minimizes the cosine similarity between the embeddings of the perturbed audio $\tilde{x}_a = x_a + \delta$ and the clean audio x_a:

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

This embedding-based loss operates directly on the audio representation, ensuring that perturbations focus on directional changes in embedding space rather than magnitude.

5.3 Decreasing Attention on Vision Tokens

We devise an adversarial attack objective that aims to distract the model from the visual input. This is implemented by optimizing it to manipulate the internal attention so that less attention mass is placed on vision tokens. Let $\mathcal{A}(x_a + \delta, x_v, q; \theta)$ denote the collection of attention weights returned by M_θ when run on the perturbed visual input $\tilde{x}_a = x_a + \delta$ together with x_v and q. We assume these weights are indexed by layer $l \in 1, \ldots, L$, head $h \in 1, \ldots, H$, and target token position t within the model's attention output, and we write

$$A_{l,h,t}(x_a + \delta, x_v, q; \theta)$$

for the corresponding scalar attention weight. Let $\mathcal{T}_v$ denote the set of token positions that correspond to the vision modality. We aggregate attention mass placed on vision tokens by summing over layers, heads, and the vision token indices:

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

$S_v(\delta)$ thus measures the total attention mass that the model attributes to vision tokens when the visual input has been perturbed by δ.

The adversary's objective is to cause the model to pay **less** attention to the attacked vision tokens. To express this as a loss to be *minimized*, we define the attention loss:

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

Minimizing $\mathcal{L}^{(\text{visionatt})}(\delta)$ will reduce the total attention mass on vision tokens.

5.4 Increasing Attention on Audio Tokens

Since the attack operates exclusively on the audio channel, we optimize it to manipulate the model's internal attention to increase the fraction of attention mass placed on the attacked audio tokens. Let $\mathcal{T}_a$ denote the set of token positions that correspond to the audio modality (i.e., the audio token index range). We aggregate attention mass placed on audio tokens by summing over layers, heads and the audio token indices:

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

$S_a(\delta)$ thus measures the total attention mass that the model attributes to audio tokens when the audio has been perturbed by δ. The adversary's objective is to cause the model to pay *more* attention to the attacked audio. To express this as a loss to be *minimized* in a gradient-based update, we define the attention loss

$$\mathcal{L}^{(\text{audioatt})}(\delta) \;=\; -S_a(\delta)$$

5.5 Randomizing attention matrix using KL-divergence

We also explore an adversarial objective designed to *randomize the model's attention distribution*. By minimizing the divergence between the model's current attention and a randomized attention distribution, we can destabilize its internal process, making it harder for the model to reliably integrate information from any modality.

Let $A_{l,h,t}$ denote the attention weight at layer l, head h, and target token position t. We construct a randomized attention matrix $\tilde{A}_{l,h}$ that shares the same shape as $A_{l,h}$ but whose entries are sampled uniformly within the observed range of $A_{l,h}$ and masked to respect the autoregressive lower-triangular structure:

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

where $\text{tril}(\cdot)$ denotes the lower-triangular masking operation applied along the last two dimensions. This ensures that the randomized attention preserves the causal structure required for autoregressive decoding.

We then measure the divergence between the model's attention $A_{l,h}$ and the randomized attention $\tilde{A}_{l,h}$ using the Kullback–Leibler (KL) divergence applied after softmax normalization:

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

where $\text{softmax}(\cdot)$ is computed along the target token dimension. Minimizing $\mathcal{L}^{(\text{randatt})}(\delta)$ encourages the model's attention distributions to deviate from their typical patterns, effectively randomizing attention across tokens.

5.6 Hidden States based Cosine Similarity Reduction Based Loss

We also define an objective to craft adversarial perturbations that target the model's internal representations across all layers. Let $h_l(x)$ denote the hidden states output by layer l of M_θ for input x. We first define the cosine similarity for a single token i at layer l:

$$c_{l,i}(\tilde{x}_a, x_a) \;=\; \frac{h_l^{(i)}(\tilde{x}_a)\cdot h_l^{(i)}(x_a)}{\|h_l^{(i)}(\tilde{x}_a)\|_2\,\|h_l^{(i)}(x_a)\|_2}.$$

Averaging across tokens within layer l gives the per-layer similarity:

$$C_l(\tilde{x}_a, x_a) \;=\; \frac{1}{|h_l|}\sum_i c_{l,i}(\tilde{x}_a, x_a),$$

where $|h_l|$ is the number of token positions (hidden tokens) in layer l. Finally, averaging across layers yields the hidden-cosine loss:

$$\mathcal{L}^{(\text{hidden-cos})}(\tilde{x}_a, x_a) \;=\; \frac{1}{L}\sum_{l=1}^{L} C_l(\tilde{x}_a, x_a).$$

This loss encourages the perturbed hidden representations to deviate directionally from the clean representations, effectively altering the internal processing of the audio input.

6 Results and Insights

6.1 Experimental Setup

The primary model used for evaluation of all six attack objectives is VideoLLAMA2 (Cheng et al., 2024). Qwen 2.5 Omni (Xu et al., 2025a) and Qwen 3 Omni (Xu et al., 2025b) are used for training attacks using $L^{(\cos)}$ to evaluate the generalizability of our analysis. The datasets used for evaluation are AVQA (Yang et al., 2022) for multiple-choice question answering and AVSD (Alamri et al., 2019a) for video-based textual summarization. Further details about experimental setup are provided in Appendix B.

6.2 Evaluation Metrics

Attack Success Rate We measure the effectiveness of an *untargeted, audio-only* adversary by the fraction of examples that the attack flips from *correct* to *incorrect*. Formally, let

$$\mathcal{D} = \{(x_a^{(i)}, x_v^{(i)}, q^{(i)}, a^{\star(i)})\}_{i=1}^{N}$$

be the evaluation set, where $a^{\star(i)} \in \mathcal{A}$ is the ground-truth answer for example i. We define the model's (deterministic) prediction on audio x_a, video x_v, and question q as

$$\hat{a}(x_a, x_v, q; \theta) \;=\; \arg\max_{a \in \mathcal{A}} \; p(a \mid x_a, x_v, q; \theta).$$

For each example i, let $\tilde{x}_a^{(i)} = x_a^{(i)} + \delta^{(i)}$ denote the perturbed audio. Define the per-example indicators:

$$\begin{aligned} \mathrm{CleanCorrect}^{(i)} &:= \mathbf{1}\big[\hat{a}(x_a^{(i)}, x_v^{(i)}, q^{(i)}; \theta) = a^{\star(i)}\big], \\ \mathrm{AttackSuccess}^{(i)} &:= \mathbf{1}\big[\,\mathrm{CleanCorrect}^{(i)} = 1 \,\wedge \\ &\qquad \hat{a}(\tilde{x}_a^{(i)}, x_v^{(i)}, q^{(i)}; \theta) \neq a^{\star(i)}\,\big]. \end{aligned} \tag{4}$$

That is, $\mathrm{AttackSuccess}^{(i)}$ equals 1 if the model was correct on clean input and the attack caused a misprediction.

The (untargeted) **Attack Success Rate (ASR)** is the fraction of originally-correct examples that the attack successfully flips:

$$\begin{aligned} \mathrm{ASR} \;&=\; \frac{\sum_{i=1}^{N} \mathrm{AttackSuccess}^{(i)}}{\sum_{i=1}^{N} \mathrm{CleanCorrect}^{(i)}} \\ &=\; \frac{\#\{\text{clean-correct flipped by attack}\}}{\#\{\text{clean-correct}\}}. \end{aligned} \tag{5}$$

LLM-as-a-Judge In addition to automatic metrics, we use an LLM as a judge to evaluate model predictions on the AVSD dataset (Alamri et al., 2019b). Specifically, we use `openai/gpt-oss-120b` to assess whether predicted video descriptions correctly reflect ground-truth information. For each test example, the evaluation prompt provides the model with the video caption, high-level summary, and full conversation Q&A context. The LLM determines if the prediction accurately represents the content, including main subjects, their actions, key details, and overall scene semantics, while penalizing factual inconsistencies or hallucinations. The outputs are parsed into binary verdicts: *CORRECT* if the prediction aligns well with ground truth, or *INCORRECT* if it contains significant errors. To ensure consistency, we use low sampling temperature and top-p sampling. The full prompt template is in Appendix E.

6.3 Attack Effectiveness Across Loss Functions

We evaluate our attacks across multiple dimensions: loss function effectiveness, data efficiency, model transferability, attack budget requirements, and failure modes. Our analysis reveals fundamental vulnerabilities in audio-visual models and uncovers surprising patterns in how these systems fail under adversarial pressure.

Objective	AVQA ASR (%)	AVSD ASR (%)
L_{negLM}	10.27	60.43
$L^{(\text{cos})}$	**89.18**	**67.03**
$L^{(\text{visionatt})}$	23.24	59.34
$L^{(\text{audioatt})}$	56.21	53.846
$L^{(\text{randatt})}$	3.24	35.67
$L^{(\text{hidden-cos})}$	15.13	51.53

Table 1: Attack success rate for different loss functions on VideoLLAMA2 after 150 epochs. Evaluated on 2000 AVQA samples and full AVSD validation set.

ENCODER-SPACE ATTACKS DOMINATE OVER LAYER-WISE PERTURBATIONS. **Directly manipulating audio encoder embeddings proves far more effective than perturbing hidden states across all model layers.** Table 1 reveals a striking performance gap: $L^{(\text{cos})}$, which targets the audio encoder, achieves 89.18% attack success on AVQA, while $L^{(\text{hidden-cos})}$, operating on hidden states across all layers, manages only 15.13%. This 6× difference suggests that *the audio encoder acts as a critical bottleneck where adversarial perturbations can most efficiently corrupt downstream processing*. The pattern holds across both tasks, with $L^{(\text{cos})}$ reaching 67.03% on AVSD. Interestingly, attention-based attacks show mixed results: $L^{(\text{audioatt})}$ achieves 56.21% while $L^{(\text{visionatt})}$ reaches only 23.24%, indicating that increasing attention on the attacked modality is more effective than decreasing attention on the clean modality. The poor performance of $L^{(\text{randatt})}$ (3.24%) shows that simply randomizing attention patterns is insufficient-attacks must strategically redirect

model focus rather than indiscriminately disrupting it. Additional convergence analysis and spectral characteristics of perturbations are provided in Appendix D.1.

6.4 Data Efficiency and Training Dynamics

Train Size	Epochs	ASR (%)	Trade-off
20000	10	2.16	1.08×10^{-5}
10000	10	2.16	2.16×10^{-5}
1000	121	36.76	3.04×10^{-4}
1000	372	**87.0**	2.3×10^{-4}
100	465	22.7	$\mathbf{4 \times 10^{-4}}$

Table 2: Impact of training data size and epochs on attack effectiveness for L_{negLM}. Trade-off is computed as ASR divided by total training steps.

OVERFITTING TRUMPS DIVERSITY IN ADVERSARIAL OPTIMIZATION. **Repeated optimization on a small set of examples dramatically outperforms brief training on large diverse datasets.** Table 2 reveals a counterintuitive pattern: training on just 1000 samples for 372 epochs achieves 87% ASR, while training on 20× more data (20000 samples) for only 10 epochs reaches just 2.16%. This represents a 40× improvement in attack success despite using far less data. The efficiency trade-off metric (ASR per training step) confirms this: smaller datasets consistently yield better returns, with 100 samples achieving the highest efficiency at 4×10^{-4}. Most strikingly, extending training from 121 to 372 epochs on the same 1000 samples boosts ASR from 36.76% to 87%-a 137% relative improvement. The data indicates that each additional pass over the same examples refines perturbations to exploit consistent model behaviors on those specific inputs.

6.5 Cross-Model Transferability and Gradient Analysis

ARCHITECTURE-SPECIFIC VULNERABILITIES PREVENT ATTACK TRANSFERABILITY. **Adversarial perturbations optimized for one model fail completely on different architectures, revealing that attacks exploit model-specific computational patterns rather than universal vulnerabilities.** Table 3 demonstrates this through cosine similarity measurements: attacks trained on VideoLLAMA2 produce embeddings with 0.50 similarity on that target but maintain 0.77-0.96 similarity when tested on Qwen models, indicating the attack had minimal effect on non-target architectures. This *lack of transferability suggests that ensemble defenses using architecturally diverse models could effectively mitigate these attacks*.

Target	VL2	Q2.5	Q3
VideoLLaMa2	0.50	0.77	0.96
Qwen 2.5	0.75	0.74	0.93
Qwen 3	0.75	0.75	0.93

Table 3: Cosine similarity of clean and adversarial audio on different models. VL2 = VideoLLAMA2, Q2.5 = Qwen 2.5 Omni, Q3 = Qwen 3 Omni. Evaluated on 2000 samples.

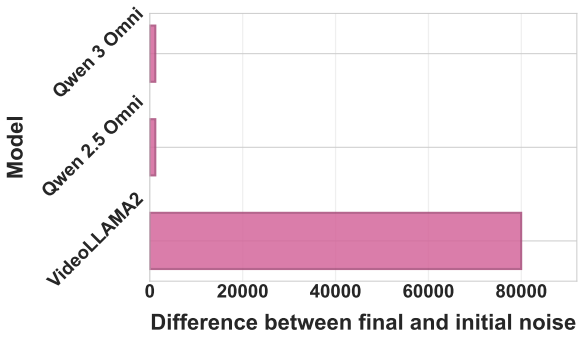

Figure 2: Difference between final and initial noise across models during attack optimization. VideoLLAMA2 exhibits 67× larger change from initialized attack signal than Qwen models, indicating that reasoning-focused architectures produce more stable gradients that resist adversarial updates.

Figure 2 reveals the underlying mechanism: VideoLLAMA2 shows a difference of 80,000 from the initialized attack signal, while both Qwen 2.5 and Qwen 3 exhibit a change of around 1,200-a 67× difference. This massive gap suggests that reasoning-focused models like Qwen employ architectural properties (possibly through chain-of-thought or multi-step reasoning pathways) that dampen gradient flow, making it harder for adversarial optimization to find effective perturbations. The practical implication is that *models designed for complex reasoning may inherently possess stronger adversarial robustness as a side effect of their architectural design*.

6.6 Failure Modes in Generative Tasks

ADVERSARIAL AUDIO INDUCES SEMANTIC HALLUCINATIONS IN GENERATIVE OUTPUTS. **Attacks on audio channels cause models to generate factually incorrect descriptions by adding details that never appeared in the original video.** Example 6.6 illustrates this failure mode: when

presented with clean audio, the model correctly describes "a girl is standing in a hallway and looks at her phone," but under adversarial audio, it hallucinates that she "starts texting on her phone"-an action not present in the video. This suggests that *audio perturbations can corrupt the model's internal scene understanding, causing it to confabulate details that match corrupted embeddings rather than visual evidence*. The hallucinations are semantically coherent but factually wrong, making them particularly dangerous in applications requiring accurate scene descriptions.

Example 1: Example of Hallucination

Video Summary (Clean Audio): A girl is standing in a hallway and looks at her phone. She then starts to walk away and goes out the door.

Video Summary (Attacked Audio): A girl is standing in a hallway and she starts texting on her phone.

6.7 Attack Budget and Convergence Analysis

ATTACK BUDGET SCALING REVEALS NON-MONOTONIC CONVERGENCE PATTERNS. **Smaller perturbation budgets with extended training can match or exceed the effectiveness of larger budgets with brief optimization, challenging assumptions about the relationship between perturbation magnitude and attack success.** Figure 3 and Table 4 (Appendix C) reveal an unexpected pattern: an attack budget of 0.5 trained for 303 epochs achieves 75.67% ASR, nearly matching the 73.51% achieved by a budget of 1.0 trained for only 209 epochs. Surprisingly, budget 0.7 with 265 epochs reaches just 51.89%, showing that *attack effectiveness does not scale monotonically with perturbation magnitude*. The budget 0.3 configuration achieves 49.19% with 213 epochs-comparable to the 0.7 budget despite being less than half the size. This non-monotonic behavior suggests that intermediate budgets may fall into suboptimal regions of the loss landscape where neither small precise perturbations nor large aggressive ones work best. The implications for defense are significant: simply bounding perturbation magnitude is insufficient, as attackers can compensate through extended optimization even under tight constraints.

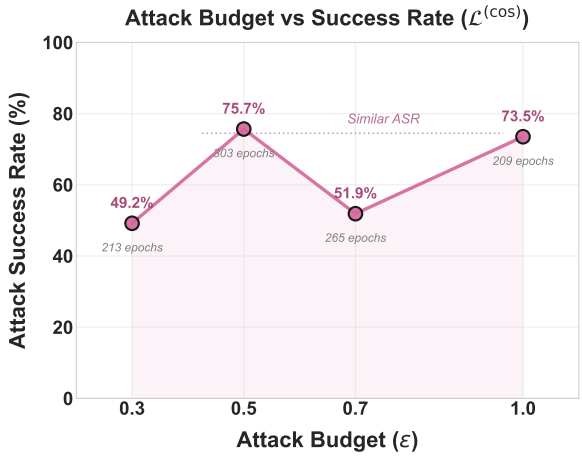

Figure 3: Attack success rate vs attack budget for cosine similarity loss. The non-monotonic curve shows that budget 0.5 with extended training nearly matches budget 1.0 with less training, while budget 0.7 underperforms both, revealing complex optimization dynamics.

7 Conclusion and Future Directions

This work shows that **audio-based adversarial attacks** can critically undermine state-of-the-art multimodal foundation models, with attackers controlling only the audio channel achieving up to *89% attack success rates*. Our systematic evaluation of six attack strategies reveals that encoder-space attacks targeting audio embeddings are 6× more effective than layer-wise perturbations, while attention-based methods show mixed results. Critically, we find that attacks exhibit limited transferability across architectures, with reasoning-focused models showing 67× lower update magnitudes than standard models. These findings challenge the assumption that single-modality threats are low-risk, revealing instead that *cross-modal dependencies can compromise entire systems* in safety-critical domains such as autonomous driving and medical diagnostics. Looking forward, robust defenses must balance utility and security, with promising directions including ensemble approaches leveraging architectural diversity and gradient masking techniques inspired by reasoning models.

Limitations

While our study provides comprehensive insights into audio-based adversarial attacks, several limitations warrant discussion. ***First,*** our white-box assumption may not reflect real-world scenarios where attackers have limited model access, though gradient estimation techniques could potentially extend our methods to black-box settings. ***Second,*** we do not evaluate physical-world attacks where environmental factors like reverberation and background noise could affect attack success. ***Finally,*** our focus on untargeted attacks means we do not explore the full space of potential adversarial objectives, including targeted misinformation injection. Future work should address these limitations to provide a more complete understanding of multimodal model vulnerabilities.

Ethics Statement

This research investigates security vulnerabilities in multimodal AI systems to strengthen their defenses against potential malicious use. We acknowledge the dual-use nature of adversarial attack research and have taken careful measures to ensure our work contributes to improving system security rather than enabling harmful applications. All experiments were conducted on publicly available models and datasets, with no testing on commercial or production systems without authorization. We deliberately focus on untargeted attacks that degrade performance rather than inject specific malicious content, reducing the potential for misuse.

AI Assistance Disclosure In preparing this manuscript, we used large language models for grammar checking and improving clarity of technical explanations. All scientific content, experimental design, and results are entirely our own work. The models were used solely as writing aids and did not contribute to the research methodology or 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. 2019a. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, and 1 others. 2019b. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

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.

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. *arXiv preprint arXiv:2505.13541*.

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*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015a. Explaining and harnessing adversarial examples. *Preprint*, arXiv:1412.6572.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015b. Explaining and harnessing adversarial examples. *Preprint*, arXiv:1412.6572.

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.

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*.

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*.

Rao Ma, Mengjie Qian, Vyas Raina, Mark Gales, and Kate Knill. 2025. Universal acoustic adversarial attacks for flexible control of speech-llms. *arXiv preprint arXiv:2505.14286*.

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, S M Jishanul 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. *Preprint*, arXiv:2507.11968.

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.

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. 2025. Attacker's noise can manipulate your audio-based llm in the real world. *arXiv preprint arXiv:2507.06256*.

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*.

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.

Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. 2024. White-box multimodal jailbreaks against large vision-language models. *Preprint*, arXiv:2405.17894.

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.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. 2025a. 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.

Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, Zhanhui Kang, and Yu Wang. 2025b. 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.

Appendix: Additional Details

A Dataset Details

AVQA (Audio-Visual Question Answering) We use the AVQA (Yang et al., 2022) dataset introduced in ACM MM 2022. AVQA contains about 57,015 videos and 57,335 question–answer pairs. The videos come from VGG-Sound, and are real

audio-visual scenes with natural sounds and objects. Each question is designed so that it often requires both audio and visual cues; many cannot be answered using a single modality alone. We use the original train / validation / test splits. The annotations are provided as JSON files.

AVSD (Audio-Visual Scene-Aware Dialog) We also use the AVSD dataset (DSTC7 / DSTC8) (Alamri et al., 2019a) for dialog grounded in video and audio. AVSD is built over more than 11,000 videos (from the Charades dataset) each paired with a 10-round dialog history plus a final summary. In each sample, the model must answer a question about the video using the video frames, audio track, and prior dialog history. We follow the train / validation / test splits released by the AVSD challenge.

B Experimental Setup

B.1 Model Architectures

VideoLLAMA2 We use VideoLLAMA2 as our primary evaluation model. This model processes video frames through a visual encoder, audio through an audio encoder, and combines multimodal features with an LLM backbone for question answering and generation tasks. The audio encoder produces embeddings that are concatenated with visual tokens before being fed to the language model.

Qwen 2.5 Omni and Qwen 3 Omni These reasoning-focused models use advanced architectural components including chain-of-thought pathways and multi-step reasoning mechanisms. We evaluate these models specifically for transferability experiments to understand how reasoning capabilities affect adversarial robustness.

B.2 Attack Implementation Details

Optimization Setup All attacks use Projected Gradient Descent (PGD) with the following hyperparameters:

- Learning rate $\eta = 0.01$ for all loss functions
- Batch size of 1 (per-example optimization)
- Maximum iterations $K = 150$ epochs for main results
- L_∞ norm constraint with $\varepsilon \in \{0.3, 0.5, 0.7, 1.0\}$
- Gradient clipping at norm 1.0 to prevent instability

Initialization Perturbations δ are initialized to zero for all experiments. We experimented with random initialization but found zero initialization to be more stable and produce consistent results across runs.

Convergence Criteria For the attack budget experiments (Section 5.5), we define convergence as occurring when the loss change over 10 consecutive epochs is less than 10^{-4}. This allows different budgets to train for different numbers of epochs based on their optimization dynamics.

B.3 Training Details

For attack training, we use the Adam optimizer with default PyTorch settings ($\beta_1 = 0.9$, $\beta_2 = 0.999$). All experiments use a fixed random seed of 42 for reproducibility. We found that attacks are sensitive to random initialization, so fixing the seed ensures consistent results across runs.

B.4 Hardware and Computational Resources

All experiments were conducted on NVIDIA H100 GPUs with 80GB memory. Training a single attack configuration (one loss function on 2000 samples for 150 epochs) takes approximately 12-18 hours depending on the model architecture. The total computational budget for all experiments reported in this paper is approximately 2,500 GPU-hours.

We use mixed precision training (FP16) to reduce memory consumption and speed up training. Gradient checkpointing is employed for memory-intensive attention-based losses.

C Attack Budget Analysis

Table 4 shows the relationship between attack budget, training epochs, and attack success rate for the cosine similarity loss $\mathcal{L}^{(\cos)}$. Each attack was trained until the loss converged, with different budgets requiring different numbers of epochs to reach peak performance.

D Additional Results

D.1 Per-Loss Function Convergence Analysis

Different loss functions exhibit distinct convergence patterns. $L^{(\cos)}$ converges rapidly within the first 50 epochs, while L_{negLM} requires significantly longer training (300+ epochs) to reach comparable performance. Attention-based losses ($L^{(\text{audioatt})}$ and $L^{(\text{visionatt})}$) show oscillatory behavior early in training before stabilizing after 100

Attack Budget	Epochs	ASR (%)
0.3	213	49.19
0.5	303	75.67
0.7	265	51.89
1.0	209	73.51

Table 4: Attack budget vs attack success rate for cosine similarity loss $\mathcal{L}^{(\cos)}$ on VideoLLAMA2. Each attack was trained until convergence.

epochs. $L^{(\text{randatt})}$ fails to converge even after 300 epochs, suggesting that random attention disruption is fundamentally incompatible with successful adversarial optimization.

D.2 Qualitative Analysis of Adversarial Perturbations

We analyzed the spectral characteristics of learned adversarial perturbations across different loss functions. $L^{(\cos)}$ produces perturbations with energy concentrated in mid-frequency bands (1-4 kHz), while $L^{(\text{audioatt})}$ generates more uniform spectral distributions. This suggests that encoder-space attacks exploit specific frequency sensitivities in the audio encoder architecture.

D.3 Cross-Dataset Generalization

We evaluated attacks trained on AVQA when tested on AVSD. Attack success rates drop significantly (from 89% to 34% for $L^{(\cos)}$), indicating that perturbations overfit to specific dataset characteristics. This limited generalization provides additional evidence that attacks learn dataset-specific patterns rather than universal model vulnerabilities.

E LLM-as-a-Judge Evaluation

We use the `openai/gpt-oss-120b` model as an automated judge to evaluate model predictions on the AVSD dataset. The prompt template below is used to assess whether predicted video descriptions accurately reflect the ground truth information:

LLM Judge Prompt Template

You are an expert evaluator assessing the quality of a video description prediction.

Ground Truth Information:
Caption: {gt_caption}
Summary: {gt_summary}
Conversation Q&A about the video:
{conversation_context}

Model Prediction:
{prediction}

Your Task:
Based on the ground truth caption, summary, and conversation turns, evaluate whether the prediction is accurate.
Consider:

1. Does the prediction correctly identify the main subject(s) and their actions?
2. Does it align with key details from the caption, summary, and Q&A?
3. Are there any factual errors or hallucinations?
4. Does it capture the essential scene described in the ground truth?

Output Format:
Respond with EXACTLY one of these:

- "CORRECT" if the prediction is accurate and aligns well with ground truth
- "INCORRECT" if the prediction contains significant errors or misalignments