NEVER STOP SPEAKING: A DENIAL-OF-SERVICE ATTACK ON END-TO-END SPEECH LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many studies have shown that specially crafted inputs can induce large language models (LLMs) to generate excessively long outputs, resulting in significant computational overhead and resource consumption. While most existing denial-of-service (DoS) attacks target text-only LLMs, end-to-end (E2E) speech LLMs are rapidly emerging. Existing text-based DoS attacks primarily rely on prompt engineering, such as adversarial suffixes or semantic inducement, which exploit the discrete nature of text inputs and therefore cannot be directly transferred to continuous speech inputs. Moreover, prior studies on speech model security mainly focus on ASR or TTS systems, leaving the DoS vulnerability of E2E speech LLMs largely unexplored. To address this gap, we propose the perturbation-based DoS attack targeting E2E speech models. Instead of inducing long outputs through prompt manipulation, our method optimizes imperceptible acoustic perturbations to directly influence the model's autoregressive generation process while preserving the original input length. Specifically, we formulate the attack as a composite optimization objective that jointly suppresses EOS generation, encourages prolonged decoding, and largely preserves semantic consistency by integrating weighted EOS loss, top-k logit loss, length loss, and semantic alignment loss. To further improve stealthiness, we employ voice activity detection (VAD) to inject perturbations only into voiced regions. Extensive experiments on three open-source E2E speech LLMs demonstrate that our method achieves stable attack success rate while significantly increasing generation length and GPU resource consumption, revealing security risks in modern ALLMs.

1 INTRODUCTION

End-to-end Audio Large Language Models (E2E ALLMs) have recently emerged as a promising paradigm for human-machine interaction, enabling direct understanding and generation of speech without relying on explicit intermediate representations. With their rapid deployment in voice assistants and conversational systems, ensuring the security and reliability of these models has become increasingly important. Among various security threats, denial-of-service (DoS) attacks pose a critical risk by forcing models to generate excessively long responses, leading to unnecessary computational consumption and service degradation. However, while DoS attacks have been extensively studied in text-based LLMs, the vulnerability of E2E ALLMs remains largely unexplored.

Existing DoS attacks against large language models Liu et al. (2026); Chen et al. (2026); Si et al. (2025) have been extensively studied. For example, adversarial suffix-based methods manipulate model behavior by appending carefully designed tokens, while recent approaches construct complex queries or decompose instructions into fine-grained reasoning steps to induce unnecessarily long generation. However, these strategies cannot be directly applied to end-to-end audio large language models (E2E ALLMs). Unlike text models that operate on editable discrete tokens, E2E ALLMs process continuous acoustic waveforms, making token-level manipulation infeasible. Moreover, prompt-based strategies typically rely on abnormally complex or unnatural inputs, which increase attack detectability and are difficult to transfer to speech inputs.Existing studies on ASR models Haque et al. (2023) have explored DoS attacks by suppressing the logits of the EOS token. However, these methods mainly target the speech-to-text component and cannot address the complex generation process of E2E ALLMs. Furthermore, optimizing only EOS logits may lack sufficient constraints on other aspects of generation, potentially degrading semantic consistency or causing

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

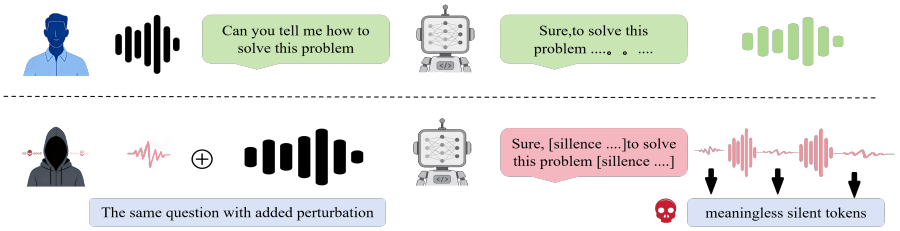

Figure 1: The figure shows the outputs of the model with and without added noise for the same input. A clean input x (top) yields a normal speech response, while the adversarial example $x + \delta$ (bottom), crafted with a tiny perturbation, induces the model to produce excessive meaningless silent segments.

unstable decoding behavior. Therefore, crafting effective DoS attacks against E2E ALLMs remains challenging for three reasons. First, perturbations must be optimized in the continuous waveform space while remaining imperceptible and preserving the original input length. Second, directly optimizing discrete generation length is non-differentiable. Third, excessive suppression of termination signals may disrupt semantic consistency or generation stability, resulting in premature decoding failures rather than prolonged outputs.

To address these issues, we propose an improved method based on multi-loss EOS logit suppression. Our loss function consists of the following four components: (1) Weighted EOS logit loss: On top of the basic summation of all EOS logits, we assign dynamic weights to the logits at each autoregressive step, incorporating token-wise penalties and generation process penalties. (2) top-k logit loss: We indirectly suppress the EOS output by increasing the probability of the top-k tokens output by the model. (3) Length loss: We force the model's output length to align with max new tokens. (4) Semantic similarity loss: We maintain the semantic consistency between the original audio and the adversarial audio to ensure attack stealthiness. We show an example in Figure 1.

To enhance attack stealthiness, we further employ voice activity detection (VAD) to extract voiced segments of the audio and add noise only to these voiced portions. Experimental results demonstrate that our method achieves strong performance on ALLM (Audio Large Language Model). Our contributions are as follows:

- We present a novel end-to-end Denial-of-Service attacks on ALLMs, filling a critical gap in end-to-end ALLM security.
- We propose a multi-objective optimization loss function to instantiate these attacks, providing a principled method for crafting adversarial inputs that disrupt ALLM behavior.
- Extensive experiments demonstrate the effectiveness of our attacks, opening new research directions and highlighting practical vulnerabilities in ALLM systems.

2 Related Work

2.1 Denial-of-service attack

The goal of Denial of Service (DoS) attacks is typically to break the output constraint of a model by constructing specially crafted inputs—such as adding adversarial suffixes, introducing spelling errors, or using homophones—thereby forcing the model to generate output indefinitely. As we know, each output token consumes substantial computational resources. Once a model begins generating unbounded responses, it continuously consumes server resources on the one hand; on the other hand, server resources are finite—once depleted, the server can no longer serve legitimate users. Previous research has demonstrated the effectiveness of DoS attacks. For example, Zhang et al. proposed the AutoDoSZhang et al. (2024) algorithm, which constructs a DoS attack tree and expands node coverage to achieve effectiveness under black-box settings. The P-DoSGao et al. (2024) attack injects poisoned samples during model fine-tuning to reduce the logits of the EOS token, thereby breaking the model's output length limit. However, due to the discrete nature of text-based methods, they

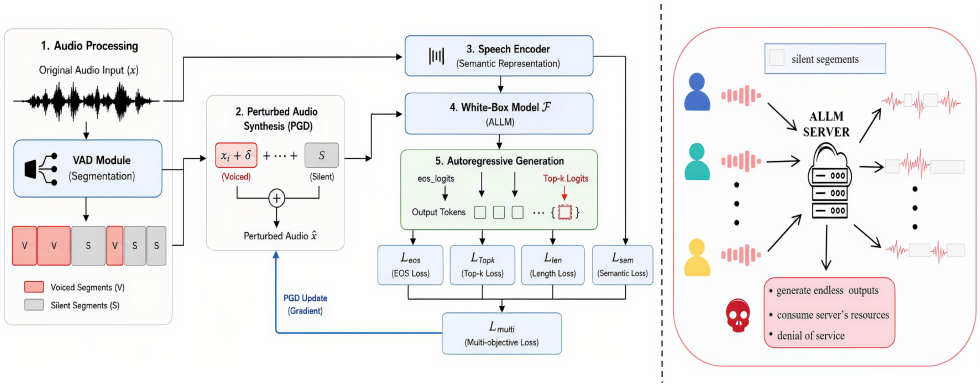

Figure 2: This figure illustrates the workflow of the adversarial DoS attack targeting end-to-end speech large models. VAD is used to separate voiced and silent segments and generate a mask for adding imperceptible perturbations. Adversarial examples are synthesized by applying tiny perturbations to the original speech. In the white-box optimization loop, we construct the multi-loss function and optimize perturbations via PGD.

cannot be directly transferred to the audio domain. Existing attacks on audio models use overly simplistic loss functions and have not achieved satisfactory performance.

2.2 E2E ALLM

The development of the Transformer has brought transformative changes to the field of artificial intelligence. With its unique attention mechanism and contextual processing capability, it has rapidly driven the evolution of LLM architectures, demonstrating superiority across various domains. Leveraging the powerful capabilities of LLMs, many studies have explored multimodal Latif et al. (2023) fusion architectures for text–video, text–image, and text–speech. Benefiting from this, speech models have gradually evolved from discrete cascade architectures (where speech input must pass through a text-based intermediate medium to produce output) to end-to-end (E2E) architectures centered around the Transformer. E2E ALLMs (Audio Large Language Models) have no explicit speech-to-text process internally Dao et al. (2024). Instead, they achieve bidirectional understanding of text and speech through text–speech alignment strategies, often employing interleaved text and speech generation strategies, demonstrating powerful speech processing capabilities.

2.3 VAD

Voice Activity Detection aims to accurately distinguish voiced segments from silence or background noise in an audio signal, and is one of the fundamental techniques in speech processing. Traditional VAD methods are mostly based on energy thresholds, zero-crossing rates, or statistical models Atal & Rabiner (1976); web (2024), but their performance is limited under low signal-to-noise ratio conditions. With the development of deep learning, neural network based VAD methods Wilkinson & Niesler (2021); Zhang et al. (2016) have significantly improved detection robustness. In the field of adversarial attacks, VAD has been used to constrain the region where perturbations are added, for instance, adding noise only to voice activity regions to improve stealthiness while maintaining attack effectiveness. Our work adopts this idea by using VAD to extract voiced segments from the audio and adding optimized perturbations only on those segments, thereby avoiding noticeable artifacts in silent regions.

3 METHOD

In this work, we build on previous research and propose an improved method that constructs a composite loss function to achieve more precise and effective attacks. To enhance stealthiness, we employ VAD to segment the audio and add noise only to the voiced portionsKo et al. (2026). Our optimization method uses gradient descent.

3.1 THREAT MODEL

We consider a denial-of-service (DoS) threat against end-to-end ALLMs. The attacker's goal is to degrade service availability by inducing the model to generate excessively long responses containing meaningless silent segments, thereby exhausting computational resources and increasing inference costs.

We focus on the white-box attack setting, where the attacker has full access to the target model architecture and parameters and can compute gradients with respect to the input speech. However, the attacker can only introduce small, imperceptible adversarial perturbations to the input audio without modifying the model or its deployment environment. The perturbation magnitude is constrained below the human auditory perception threshold to ensure stealthiness.

3.2 ATTACK OVERVIEW

Prior work has shown that large language models (LLMs)OpenAI (2023) are susceptible to denial-of-service (DoS) attacks. Conventional DoS methods targeting text-based models typically append adversarial suffixes to the input text—a technique that introduces discrete tokens and is thus inapplicable to the continuous audio modality. Alternative approaches craft complex semantic constructs or sub-queries to elicit longer outputs; however, naively converting such crafted text into speech and feeding it to the model lacks novelty, and our subsequent experiments confirm that this strategy is largely ineffective.

Therefore, we propose our method. We consider a white-box attack scenario, where the attacker has full access to the architecture and parameters of the target end-to-end audio large language model (E2E ALLM) and can compute the gradients of the model's output. Given an initial audio x, a white-box model F to be attacked, and a composite loss function L, we first apply VAD to segment x into voiced and silent segments. We extract all voiced segments $V = \{x_1, x_2, \ldots, x_n\}$,and silent segments $S = \{y_1, y_2, \ldots, y_n\}$. The goal of our attack is to generate noise $N = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ for each voiced segment such that adding these noises to the original speech forces the model F to output as long as possible. It can be formulated as a constrained optimization problem:

$$\min_{\{\sigma_1,\sigma_2,\ldots,\sigma_n\}} \quad \mathcal{L}\left(F(\hat{x})\right), \tag{1}$$

$$s.t. \quad \|\sigma_i\|_p \leq \epsilon_i, \quad \forall i \in \{1, 2, \ldots, n\}, \tag{2}$$

$$\hat{x} = Concat\left(\{x_i + \sigma_i\}_{i=1}^{n},\ S\right) \tag{3}$$

where $\mathcal{L}$ is our proposed composite loss function detailed in the next subsection. By minimizing $\mathcal{L}$, the optimized perturbation effectively suppresses the End-of-Sentence (EOS) token probability while preserving semantic coherence, leading to prolonged autoregressive generation. Figure 2 illustrates the overall workflow of our attack.

3.3 NOISE OPTIMIZATION BASED ON COMPOSITE LOSS FUNCTION

Optimization Pipeline Given the initial audio $\mathbf{x}$ and the target model F, we iteratively optimize the perturbation via gradient descent. At the beginning of each iteration, we concatenate the perturbed voiced segments (initialized with random Gaussian noise) with the unaltered silent segments to reconstruct the full adversarial waveform $\hat{x}$. We then feed $\hat{x}$ into the speech encoder of the target model to extract acoustic features, followed by the autoregressive decoder. During the decoding process, we record three types of outputs at each step: the logit of the EOS token, the logits of the top-K tokens, and the final total generation length N. After the generation finishes (either by emitting EOS or reaching the maximum limit), we compute the composite loss $\mathcal{L}_{DoS}$ using the recorded values. The gradient of $\mathcal{L}_{DoS}$ with respect to the perturbation is calculated via backpropagation, and the perturbation is updated using Projected Gradient Descent (PGD)Madry et al. (2018). This optimization loop repeats until the generated response reaches the preset maximum token limit $N_{\max}$ or the maximum iteration count is exhausted.

The overall composite loss is defined as:

$$\mathcal{L}_{DoS} = \mathcal{L}_{eos} + \mathcal{L}_{topk} + \mathcal{L}_{len} + \mathcal{L}_{sem} \tag{4}$$

In this loss function, the terms can be mainly divided into direct losses and indirect losses. The direct losses focus on controlling the generation length of the model by suppressing the EOS logits and stretching the length of the speech tokens output by the model. The indirect losses achieve indirect suppression of the EOS logits and smooth the model behavior by increasing the probability of top-k token outputs and preserving semantic invariance of the noisy speech, thereby raising the probabilities of other tokens and preventing premature termination caused by semantic issues.

Direct losses. The direct losses consist of two parts. The first is the EOS logits loss. Unlike the traditional approach of simply summing all EOS logits, we assign a weight to each EOS logit as follows:

$$\mathcal{L}_{eos} = \frac{1}{N} \sum_{t=1}^{N} w_t \ell_{EOS,t} \tag{5}$$

The weight consists of a sign-based gating component and a dynamically growing component. The gating component depends on the sign of the logit value. Since a non-positive EOS logit implies that the EOS token is almost impossible to be output, we set its weight to zero to completely exclude such uninformative steps from the optimization. Only steps with positive EOS logits are assigned a positive weight, directing the optimizer toward positions where EOS suppression is actually meaningful. The dynamic component depends on the depth of the autoregressive step. It is known that the probability of outputting the EOS token increases as generation proceeds. Therefore, this component gradually increases with the autoregressive step, and we adopt exponential growth to make the optimization pay more attention to later stages of the model output.

$$w_t = \begin{cases} w_{high} \exp\left(\frac{t}{N}\kappa\right), & \ell_{EOS,t} > 0 \\ 0, & \ell_{EOS,t} \leq 0 \end{cases}$$

To encourage long generation sequences, we introduce a differentiable length loss. Since the generated audio token length is determined by the EOS prediction during autoregressive decoding, directly optimizing the discrete output length is not feasible. Therefore, we estimate the expected generation length based on the probability distribution of EOS termination at each generation step. Let z_t denote the EOS logit at the t-th generation step and $p_t = \sigma(z_t)$ be the corresponding EOS probability. The probability that the generation terminates at step t is computed by considering that no EOS token is generated before step t and an EOS token is generated at step t. The expected generation length is then formulated as:

$$\mathbb{E}[L] = \sum_{t=1}^{N} t \left(p_t \prod_{i=1}^{t-1} (1 - p_i) \right) \tag{6}$$

where N is the maximum generation length. The length loss is defined as:

$$\mathcal{L}_{Len} = \frac{(N_{\max} - \mathbb{E}[L])^2}{N_{\max}}. \tag{7}$$

This formulation encourages the expected generation length to approach the maximum allowed length while maintaining a comparable scale with other loss terms.

Indirect losses. The indirect losses also consist of two parts. The first is the top-k loss. At each generation step, the model samples the output from the top-k tokens. Therefore, the top-k loss focuses on increasing the logits of the top-k tokens. The increase in top-k logits indirectly reduces the EOS logits. At the same time, increasing the top-k logits helps preserve the coherence of the synthesized audio.

$$\mathcal{L}_{topk} = -\frac{1}{N} \sum_{t=1}^{N} \frac{1}{K} \sum_{k=1}^{K} \ell_{topk,t}^{(k)} \tag{8}$$

Finally, we introduce a semantic alignment loss. For the original speech and the noise-added speech, we extract their features using a speech encoder and compute the cosine similarity between the two

Table 1: Attack performance comparison across different models and baselines. All experiments are performed on the OpenSLR and QCRI datasets across evaluated models.

Model	Method	ASR (↑)	AVG. Output Length (↑)	Memory Usage (GB)
Liquid Audio (1.5B)	Clean audio	0.00	198.34	8.89/47.99
	Random Noise	0.00	205.51	8.58/47.99
	Simple Loss	0.79	857.38	10.56/47.99
	Crabs	0.34	545.72	9.96/47.99
	ExtendAttack	0.29	496.37	10.18/47.99
	Our method	**0.87**	**941.88**	10.78/47.99
FunAudioChat (8B)	Clean audio	0.00	213.57	20.26/47.99
	Random Noise	0.00	210.83	20.29/47.99
	Simple Loss	0.77	869.52	21.74/47.99
	Crabs	0.36	594.38	21.50/47.99
	ExtendAttack	0.48	607.54	21.16/47.99
	Our method	**0.84**	**920.24**	21.93/47.99

feature vectors. This ensures that the perturbation does not affect the semantics of the audio, thereby eliminating the possibility that changes in output length are caused by semantic alteration.

$$\mathcal{L}_{sem} = 1 - \frac{\langle s(x), s(x+\delta) \rangle}{\|s(x)\|_2 \|s(x+\delta)\|_2} \tag{9}$$

4 Experiment

4.1 Experiment Setup

Model and Datasets We evaluate our audio adversarial attack on three recent open-source audio large models: LFM2.5-Audio-1.5BAI (2025),Fun-Audio-Chat-8BTan et al. (2025) and Qwen2-Audio-7B-InstructChu et al. (2024). LFM2.5-Audio is a lightweight end-to-end foundation model designed for low-latency real-time speech conversation; Fun-Audio-Chat and Qwen2-Audio are medium-sized audio-language dialogue models optimized for general speech interaction and audio analysis. We evaluate on two primary models and further report Qwen2-Audio results in Appendix C. For evaluation, we use two public speech datasets, OpenSLR and QCRI.

Attack Setup During the attack phase, we adopt the Projected Gradient Descent (PGD) framework to optimize adversarial perturbations, aiming to induce the target model to generate longer outputs. Specifically, we randomly sample 100 clean audio samples from each dataset, add the adversarial perturbations (generated by our method) to the original audio, and feed the perturbed audio into the target models to measure the attack success rate. Due to limited computational resources, we set $N_{\max} = 1024$.

The hyperparameters are set as follows. For $\mathcal{L}_{eos}$, we set the $w_{high} = 4$ and the exponential growth parameter $\kappa = 4$. For $\mathcal{L}_{topk}$, we set $k = 3$. During optimization, we run PGD for 200 iterations per sample, with the perturbation magnitude constrained by an ℓ_∞ norm bound of $\epsilon = 10^{-4}$.

Metrics To comprehensively evaluate the effectiveness and resource cost of our attack method across different baselines and all reported results are averaged over 100 samples. We adopt the following three metrics:

- **Attack Success Rate (ASR):** The proportion of adversarial samples that successfully force the target model to generate audio token sequences reaching the maximum generation length limit.
- **Output Token Length:** The number of tokens in the model's generated response. We report the average output length over all test samples, which helps analyze how the attack influences the model's generative behavior.
- **Peak GPU Memory Usage:** The maximum GPU memory allocated during a single forward inference pass (without gradient computation), measured in MB or GB. This metric indicates the practical computational overhead and feasibility of the attack, especially for edge-side audio models.

Table 2: Ablation study on loss components $\mathcal{L}_{sem}$, $\mathcal{L}_{len}$ (length loss), $\mathcal{L}_{eos}$, and $\mathcal{L}_{topk}$.

$\mathcal{L}_{sem}$	$\mathcal{L}_{len}$	$\mathcal{L}_{eos}$	$\mathcal{L}_{topk}$	VAD	ASR (↑)	Avg.Output Length (↑)	Response Quality (↑)
✓	✓	✓		✓	0.80	893.47	4.46
✓	✓		✓	✓	0.12	289.43	4.71
✓		✓	✓	✓	0.79	867.65	4.70
	✓	✓	✓	✓	0.83	937.81	3.92
✓	✓	✓	✓		0.84	945.33	4.52
✓	✓	✓	✓	✓	0.84	950.24	4.75

- **Response Quality:** We extract the generated text tokens from clean and adversarial inputs and use ChatGPT 5.5OpenAI (2025) as an evaluator to assess semantic consistency and response quality. The final score is averaged over all samples on a 1-5 scale. The evaluation prompt is provided in Appendix A.

To demonstrate the effectiveness of our proposed attack method, we compare it with the following three baselines: (1) using clean original audio or audio with random noise as input; (2) a conventional method that relies solely on naive eos logits; (3) a noise injection approach that operates on the entire global audio without employing Voice Activity Detection (VAD). (4) CrabsZhang et al. (2024), a kind of DoS attack on natural LLM (5) ExtendAttackZhu et al. (2025),a kind of dos attack on LRM

4.2 Main Result

Table 1summarizes the attack performance under the default top-k sampling strategy. Our method consistently achieves the strongest DoS capability across different E2E ALLMs, demonstrating its effectiveness in manipulating autoregressive generation behaviors. On LFM2.5-Audio and FunAudioChat, our method achieves attack success rates of 87% and 84%, respectively, significantly outperforming all baseline methods. Meanwhile, the generated output length is extended to 941.88 and 920.24 tokens, which is more than four times longer than clean inputs and substantially exceeds random noise and existing DoS baselines. This indicates that the optimized acoustic perturbations can effectively prevent normal termination and force the model to continue decoding for an extended period, thereby introducing considerable computational overhead.

Compared with the simple EOS suppression baseline, our method achieves higher ASR and longer outputs on both models. This improvement demonstrates that directly optimizing EOS logits alone is insufficient for reliable DoS attacks, while jointly modeling termination suppression, generation length, token distribution, and semantic consistency provides a more effective optimization objective. In addition, text-based DoS attacks, including Crabs and ExtendAttack, exhibit limited effectiveness when transferred to the audio domain, achieving much lower ASR and shorter outputs. These results further verify that discrete token manipulation strategies designed for text LLMs cannot be directly applied to continuous speech inputs.

Beyond generation length, our attack also introduces practical computational costs. As shown in Table 1, the increased output length leads to substantial GPU memory consumption during inference, with our method requiring 10.78 GB and 21.93 GB on LFM2.5-Audio and FunAudioChat, respectively, compared with only 8.89 GB and 17.26 GB for clean inputs. These results reveal that carefully optimized acoustic perturbations can effectively exploit the autoregressive generation mechanism of E2E ALLMs and cause significant resource consumption while preserving the original input semantics.

To further evaluate the robustness of our method, we additionally test different decoding strategies and the Qwen2-Audio model. We also test the attack transferability. The results under greedy decoding, on Qwen2-Audio and the transferability are reported in the Appendix C.

4.3 Output Audio Analysis

To further understand the mechanism behind output elongation, we analyze the generated waveforms and intermediate text responses from successfully attacked samples, as shown in Figure 3a.

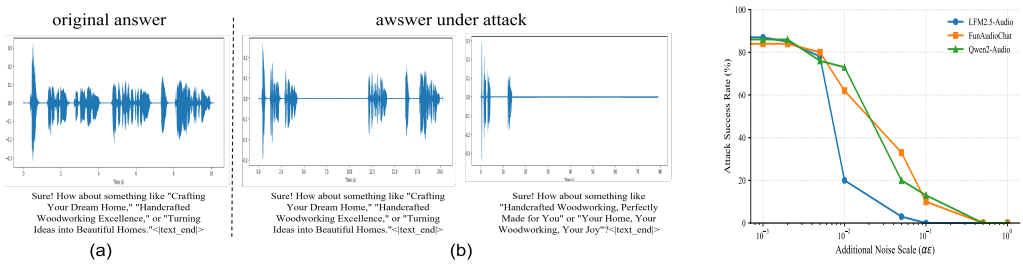

Figure 3: (a) Waveform comparison between clean output and attacked output. (b) ASR under additional noise.

The transcribed responses remain largely consistent with those generated from clean inputs, indicating that our attack does not significantly affect the model's semantic understanding capability. However, waveform analysis reveals that attacked outputs contain substantially more low-energy silent regions compared with clean outputs. This phenomenon is mainly caused by the joint optimization of L_{EOS} and L_{len}, which encourages the model to continue autoregressive decoding beyond its normal termination point. As generation extends beyond meaningful response ranges, the additional tokens tend to correspond to low-information speech segments, resulting in prolonged outputs and increased computational consumption.

4.4 Ablation Study

Effectiveness of Loss Components. To investigate the contribution of each component in our composite objective, we conduct ablation experiments by removing each loss term from L_{DoS} individually while keeping all other settings unchanged. The results are summarized in Table 2.

Removing L_{EOS} leads to a significant degradation in attack performance, reducing the ASR from 0.84 to 0.12 and the average output length from 950.24 to 289.43 tokens. This demonstrates that weighted EOS suppression is the primary factor driving prolonged generation, as it directly prevents premature termination during autoregressive decoding. Without L_{len}, the ASR decreases slightly to 0.79 and the output length drops to 867.65 tokens, indicating that the length loss provides additional optimization guidance toward the maximum generation boundary. Similarly, removing L_{topk} causes a moderate reduction in ASR and output length, while noticeably degrading response quality. This suggests that L_{topk} mainly improves generation stability by redistributing token probabilities and indirectly suppressing EOS prediction.

Removing L_{sem} slightly improves ASR and output length, but results in a substantial decline in response quality (from 4.75 to 3.92). This indicates that semantic alignment is not essential for maximizing output length, but plays a critical role in preserving the original speech semantics and ensuring attack stealthiness. Overall, the ablation results verify that each component serves a complementary purpose: L_{EOS} provides the fundamental attack capability, L_{len} and L_{topk} enhance generation extension and stability, while L_{sem} maintains semantic consistency.

Effectiveness of the VAD Strategy. To evaluate the contribution of the VAD-based perturbation strategy, we compare our full method with a variant that applies perturbations to the entire waveform without VAD. As shown in Table2, removing VAD results in a similar attack success rate (0.84) and slightly shorter output length, indicating that restricting perturbations to voiced regions does not compromise the attack effectiveness. Meanwhile, the response quality decreases from 4.75 to 4.52 without VAD, suggesting that perturbing silent regions introduces unnecessary acoustic interference and affects semantic preservation. These results demonstrate that VAD provides a better trade-off between attack effectiveness and stealthiness by concentrating perturbations on informative speech regions.

Effect of the exponential growth parameter κ. We further investigate the impact of the exponential growth parameter κ in L_{eos}. As shown in Table 3, the attack performance first improves and then decreases as κ increases, with $\kappa = 4$ achieves the best or comparable performance across all eval-

Table 3: Ablation study on the exponential growth parameter κ in L_{eos}.

Model	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$	$\kappa = 6$
LFM2.5-Audio	82%	84%	**87%**	86%	79%
FunAudioChat	80%	82%	**84%**	82%	76%
Qwen2-Audio	81%	83%	**83%**	79%	77%

uated models. When κ is relatively small, the dynamic EOS weighting grows slowly, limiting the emphasis on later autoregressive steps where EOS prediction becomes more critical. Consequently, the optimizer may fail to sufficiently suppress EOS probabilities within the fixed optimization budget. In contrast, overly large κ values cause the EOS weights to increase too rapidly, which may introduce excessively large gradients and make the optimization unstable. Moreover, an excessively dominant L_{eos} term can weaken the contributions of other loss components, such as length guidance and semantic preservation, resulting in degraded attack effectiveness. Therefore, we select $\kappa = 4$ as it provides a balanced trade-off between effective EOS suppression and stable multi-objective optimization.

4.5 Attack Robustness

We further evaluate the robustness of our adversarial perturbations against additional random noise. Specifically, after generating adversarial examples, we add extra noise η with different magnitudes relative to the original perturbation constraint ϵ, i.e., $\|\eta\|_\infty = \alpha\epsilon$, and measure the resulting ASR.

As shown in Fig. 3b, our attack remains effective under small additional perturbations, indicating that the optimized perturbations can tolerate moderate input variations. It is worth noting that $\alpha\epsilon$ represents the noise scale relative to the predefined perturbation bound rather than the actual magnitude of the optimized perturbation. Since the final perturbations generated by PGD may occupy different magnitude ranges across different models, the point at which additional noise starts to disrupt the attack varies among models. When the added noise reaches a comparable magnitude level to the optimized perturbation, the attack effectiveness gradually decreases, as the original adversarial direction is disturbed.

5 Conclusion

In this paper, we present a systematic study of denial-of-service (DoS) attacks against end-to-end audio large language models (E2E ALLMs). We propose a white-box adversarial attack framework that leverages VAD-based perturbation and a multi-component optimization objective to suppress EOS prediction and induce prolonged autoregressive generation. Extensive experiments on multiple E2E ALLMs demonstrate that the proposed method can effectively increase output lengths, achieve high attack success rates, and introduce substantial inference overhead while largely preserving the semantic information of the original speech.

Through comprehensive ablation studies, we analyze the contribution of each loss component and show that weighted EOS suppression, length guidance, token distribution optimization, and semantic alignment jointly improve attack effectiveness and stealthiness. Furthermore, our robustness analysis reveals that the attack can tolerate moderate input perturbations. These findings highlight the potential security risks of current E2E ALLMs and motivate the development of practical defense mechanisms against audio-domain adversarial DoS attacks.

REFERENCES

Webrtc voice activity detector. https://webrtc.org/, 2024. Open-source real-time voice activity detection algorithm widely adopted in speech processing pipelines.

Liquid AI. Lfm2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

B. Atal and L. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):201–212, 1976. doi: 10.1109/TASSP.1976.1162800.

Yiming Chen, Zexin Li, Xianghu Yue, Robby T. Tan, and Haizhou Li. NaturalSloth: Revisiting denial-of-service attacks on large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19685–19702, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.901. URL https://aclanthology.org/2026.acl-long.901/.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Alan Dao, Dinh Bach Vu, and Huy Hoang Ha. Ichigo: Mixed-modal early-fusion realtime voice assistant, 2024.

Kuofeng Gao, Tianyu Pang, Chao Du, Yong Yang, Shu-Tao Xia, and Min Lin. Denial-of-service poisoning attacks against large language models, 2024.

Mirazul Haque, Rutvij Shah, Simin Chen, Berrak Şişman, Cong Liu, and Wei Yang. Slothspeech: Denial-of-service attack against speech recognition models, 2023.

Kyoungmin Ko, Sunghwan Kim, and Hyun Kwon. Audio adversarial example with no noise in the silent area for speech recognition system. *IEEE Access*, 14:2924–2938, 2026. doi: 10.1109/ACCESS.2025.3598087.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuay'ahuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, and Björn W. Schuller. Sparks of large audio models: A survey and outlook, 2023.

Xiaogeng Liu, Xinyan Wang, Yechao Zhang, Sanjay Kariyappa, Chong Xiang, Muhao Chen, G. Edward Suh, and Chaowei Xiao. Reasoningbomb: A stealthy denial-of-service attack by inducing pathologically long reasoning in large reasoning models, 2026.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

OpenAI. Periodic outages across ChatGPT and API, 2023. URL https://status.openai.com/incidents/21vl32gvx3hb. Accessed on 16/03/2024.

OpenAI. Gpt-5.5 technical report. OpenAI Technical Report, 2025. URL https://openai.com/.

Wai Man Si, Mingjie Li, Michael Backes, and Yang Zhang. Excessive reasoning attack on reasoning llms, 2025. URL https://arxiv.org/abs/2506.14374v1.

Chao-Hong Tan, Qian Chen, Wen Wang, Chong Deng, Qinglin Zhang, Luyao Cheng, Hai Yu, Xin Zhang, Xiang Lv, Tianyu Zhao, Chong Zhang, Yukun Ma, Yafeng Chen, Hui Wang, Jiaqing Liu, Xiangang Li, and Jieping Ye. Drvoice: Parallel speech-text voice conversation model via dual-resolution speech representations, 2025. URL https://arxiv.org/abs/2506.09349.

Nicholas Wilkinson and Thomas Niesler. A hybrid cnn-bilstm voice activity detector. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6803–6807, 2021. doi: 10.1109/ICASSP39728.2021.9415081.

Yu Zhang, Yuxuan Wang, and James Glass. Feature learning with raw-waveform cldnns for voice activity detection. In *Interspeech*, pp. 3542–3546, 2016.

Yuanhe Zhang, Zhenhong Zhou, Wei Zhang, Xinyue Wang, Xiaojun Jia, Yang Liu, and Sen Su. Crabs: Consuming resource via auto-generation for llm-dos attack under black-box settings. *arXiv preprint arXiv:2412.13879*, 2024. arXiv:2412.13879v4 [cs.CL].

Zhenhao Zhu, Yue Liu, Zhiwei Xu, Yingwei Ma, Hongcheng Gao, Nuo Chen, Yanpei Guo, Wenjie Qu, Huiying Xu, Zifeng Kang, Xinzhong Zhu, and Jiaheng Zhang. Extendattack: Attacking servers of lrms via extending reasoning, 2025. URL https://arxiv.org/abs/2506.13737.

A Response Quality Evaluation Prompt

To evaluate response quality, we use an LLM-as-a-Judge strategy. The generated text responses from clean and adversarial inputs are provided to ChatGPT for evaluation. Each response is assigned a score from 1 to 5, and the final Response Quality score is calculated by averaging the scores over all test samples.

Prompt for Response Quality Evaluation

Instruction: You are an evaluator for speech language model responses.
Given a reference response generated from clean audio and an adversarial response generated from perturbed audio, evaluate whether the adversarial response preserves the semantic information and task completion ability of the reference response.
Please assign an integer score from 1 to 5:
5: The adversarial response fully preserves the semantic content and quality of the reference response.
4: The adversarial response is mostly consistent with the reference response with minor degradation.
3: The adversarial response preserves the main information but has noticeable degradation.
2: The adversarial response has significant semantic differences or missing information.
1: The adversarial response fails to preserve the original meaning or task requirement.
Only output the score.

Figure 4: Prompt used for LLM-based response quality evaluation.

B Attack Performance under Greedy Decoding

In the main experiments, we adopt top-k sampling as the default decoding strategy. To investigate whether the effectiveness of our attack depends on stochastic sampling behaviors, we further evaluate our method under greedy decoding, where the token with the highest probability is selected at each generation step.

As shown in Table 4, our method maintains strong attack performance under deterministic decoding. Specifically, it achieves ASRs of 86%, 84%, and 85% on Liquid Audio, FunAudioChat, and Qwen2-Audio, respectively, while extending the average output length to over 900 tokens across all evaluated models. These results demonstrate that the proposed perturbations do not rely on sampling randomness, but instead directly influence the autoregressive generation process by suppressing termination and encouraging continuous decoding.

The consistent performance between top-k sampling and greedy decoding further verifies the robustness of our optimization objective, showing that the attack can generalize across different decoding strategies and effectively expose the inherent generation-control vulnerability of E2E ALLMs.

C Evaluation under Qwen2-Audio

To further evaluate the generality of our attack across different E2E ALLM architectures, we conduct additional experiments on Qwen2-Audio-7B-Instruct. As shown in Table 5, our method achieves an

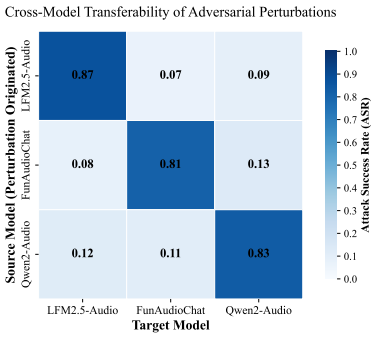

Figure 5: attack transferability

Table 4: Attack performance under greedy decoding strategies

Model	Method	ASR (↑)	AVG. Output Length (↑)	Memory Usage (GB)
Liquid Audio (1.5B)	Clean audio	0.00	162.52	8.07/47.99
	Random Noise	0.00	161.28	8.96/47.99
	Simple Loss	0.77	832.58	10.46/47.99
	Crabs	0.34	538.24	9.92/47.99
	Our method	**0.86**	**927.69**	10.27/47.99
FunAudioChat (8B)	Clean audio	0.00	187.22	20.53/47.99
	Random Noise	0.00	195.79	20.76/47.99
	Simple Loss	0.76	852.34	21.19/47.99
	Crabs	0.38	613.72	21.07/47.99
	Our method	**0.84**	**905.36**	21.44/47.99
QwenAudio(7B)	Clean audio	0.00	182.35	18.37/47.99
	Random Noise	0.00	181.43	18.12/47.99
	Simple Loss	0.74	807.75	19.09/47.99
	Crabs	0.43	643.35	20.07/47.99
	Our method	**0.85**	**934.02**	20.14/47.99

ASR of 83% and increases the average output length to 913.07 tokens, significantly outperforming clean inputs and random noise baselines. Meanwhile, the generated responses consume 19.94 GB GPU memory during inference, demonstrating that the optimized perturbations can effectively induce prolonged decoding and introduce additional computational overhead.

Compared with the simple EOS suppression baseline, our method achieves higher attack effectiveness and longer generated sequences. This indicates that relying solely on EOS logit optimization is insufficient for stable DoS attacks, while jointly optimizing termination suppression, generation length, token distribution, and semantic consistency provides a more reliable attack objective. The consistent performance across Qwen2-Audio and the other evaluated E2E ALLMs demonstrates that our approach can effectively exploit the shared autoregressive generation mechanism of different audio language models.

Table 5: Attack performance on Qwen2-Audio

Model	Method	ASR (↑)	AVG. Output Length (↑)	Memory Usage (GB)
Qwen-Audio(7B)	Clean audio	0.00	187.13	18.42/47.99
	Random Noise	0.00	179.46	18.37/47.99
	Simple Loss	0.80	846.53	20.01/47.99
	Crabs	0.42	665.19	19.07/47.99
	Our method	**0.83**	**913.07**	19.94/47.99

D ATTACK TRANSFERABILITY

To evaluate the black-box transferability of our attack, we generate adversarial perturbations on one source model and directly apply them to other unseen target models. As shown in Fig5, the diagonal

entries represent white-box attacks, achieving ASRs of 87%, 81%, and 83% on LFM2.5-Audio, FunAudioChat, and Qwen2-Audio, respectively. For cross-model transfer, the perturbations achieve non-zero ASRs across all settings, with transfer rates ranging from 7% to 13%. Models with similar scales exhibit relatively higher transferability, indicating that the optimized perturbations capture partially shared generation characteristics among E2E audio models. These results demonstrate the feasibility of our method in practical black-box scenarios.

E AI USE STATEMENT

We leveraged ChatGPT for language refinement, draft writing and code documentation assistance. The authors hold full responsibility for the paper, and all experiments, code, results, citations and scientific statements were independently reviewed and confirmed by the authors.

F REPRODUCIBILITY STATEMENT

Section 4.1 describes datasets, target models, hyper-parameters, and evaluation metrics. We sample 100 audio examples from OpenSLR and QCRI datasets and test on three open-source E2E audio large-language models. The full prompt for LLM-based response-quality evaluation is included in Appendix A. Our supplementary package provides experiment logs and plotting scripts for reproducing tables and figures. Source code will be released after acceptance. All reported numbers are averaged over test samples. Ablation experiments on loss terms, VAD masking, hyper-parameter κ, decoding strategies and transferability validate our design decisions.