Speech Tokenizers are Vulnerable: Transferable Semantic Attack and Robust Tokenizer

Zhisheng Zhang[1†], **Yifan Mi**[2†], **Yixuan Zhou**[1], **Haiyun Li**[1], **Shengbo Cai**[1], **Dongfu Song**[1], **Ziang Xu**[1], **Jie Hao**[2‡], **Zhiyong Wu**[1‡]
[1] Shenzhen International Graduate School, Tsinghua University
[2] Beijing University of Posts and Telecommunications
zhangzs25@mails.tsinghua.edu.cn

Abstract

Speech tokenizers serve as speech front ends, encoding input speech signals into representations for downstream models, such as automatic speech recognition (ASR) systems and large audio-language models (LALMs). In this paper, we find that speech tokenizers are fragile at the semantic level: small perturbations can disrupt their semantic encoding process and consequently cause severe degradation in downstream ASR and LALM performance. To study this vulnerability, we propose **T-SemAttack**, a transferable semantic attack based on a semantic encoder ensemble. By jointly drifting the representation spaces of multiple semantic encoders, T-SemAttack effectively disrupts the semantic content of speech while preserving perceptual quality, thereby inducing error features. We further analyze layer-wise representation drift and cross-model gradient alignment. Our findings show that perturbations can induce substantial shifts in deep semantic representations and that the input gradients of different surrogate encoders exhibit directional alignment. These observations suggest that shared adversarially sensitive directions in the input space may be a key factor underlying cross-model attack transferability. Building on these, we introduce **ROSETok**, a robust speech tokenizer that combines noisy training with robust semantic distillation to improve reconstruction fidelity and downstream SpeechLLM robustness. Extensive experiments on multiple datasets, 9 open-source and black-box ASR systems, and 4 LALMs demonstrate that T-SemAttack achieves strong transferable attack performance, while the proposed Robust Speech Tokenizer exhibits robustness under high-fidelity reconstruction and downstream tasks. Our code and demo are available at https://wxzyd123.github.io/t-semattack.

1 Introduction

In recent years, LALMs [1, 2] have demonstrated exceptional interaction capabilities and are widely deployed in scenarios, e.g., in-vehicle assistants and meeting transcription. Compared to previous speech models [3, 4], current large language model (LLM)-based architectures [1, 5, 6] typically rely on a speech tokenizer to map continuous signals into discrete tokens, which are subsequently fed into LLMs for modeling. Consequently, the speech tokenizer bridges the continuous and discrete representations, and its robustness directly dictates the reliability of the entire semantic pipeline.

Existing studies primarily assess the vulnerability of speech tokenizers to Gaussian noise [7], whereas semantic attacks can expose more severe vulnerabilities. As shown in Figure 1, the semantic attack (T-SemAttack) causes substantially greater degradation in downstream FSC performance across

[†]Equal contribution.
[‡]Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

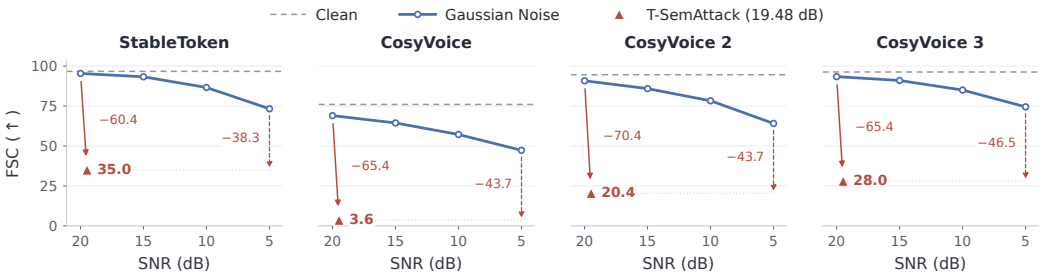


Figure 1: Gaussian Noise vs. T-SemAttack across Tokenizers. We feed the pre-quantized representations from each tokenizer into **SmolLM2-135M** [10] and evaluate downstream **semantic understanding on the FSC spoken command intent recognition task (↑)**.

models than Gaussian noise, yielding scores even lower than those obtained under stronger Gaussian noise at an SNR of 5 dB. These results indicate that tolerance to random noise does not guarantee robustness against perturbations optimized to disrupt semantic representations.

In this paper, we propose a **T**ransferable **Sem**antic **Attack** (**T-SemAttack**), which focuses representation drifting in the speech tokenizer. By jointly perturbing multiple surrogate semantic encoders [8], the method directly corrupts the semantic content of the input speech. We employ the psychoacoustic model [9] to preserve the perturbative perceptual quality. The resulting adversarial examples can therefore effectively interfere with the encoding process of the speech tokenizer while maintaining high perceptual quality, leading to severe semantic deviations in both ASR and LALM systems.

Furthermore, we analyze the effectiveness and transfer mechanisms of T-SemAttack from three perspectives: layer-wise representation drift, shared attack directions across models, and perturbation magnitude. We investigate how perturbations affect deep representations, whether shared sensitive directions underpin cross-model transferability, and how representation shifts translate into token changes and downstream recognition errors as perturbation magnitude increases. These analyses inform the training of tokenizers for semantic stability.

We consider the speech tokenizer as the safety bottleneck in ASR or LALMs. Based on our analysis of the attack, we argue that defense mechanisms should not remain only at the level of the entire model, but should also consider the speech tokenizer. To this end, we propose a **RO**bust **S**p**E**ech **Tok**enizer (**ROSETok**) that achieves high-fidelity reconstruction while preserving semantically rich representations with downstream SpeechLLM against Gaussian noise and adversarial attacks.

To evaluate the effectiveness of the T-SemAttack, we conduct experiments on bilingual and multi-speaker datasets, 9 ASR systems, and 4 LALM systems, demonstrating that T-SemAttack achieves state-of-the-art (SOTA) performance in terms of perceptual quality, attack efficiency, and effectiveness. In addition, we verify the robustness of T-SemAttack against adaptive defenses, including denoising and data transformations. Finally, we experimentally assess the reconstruction and downstream semantic capabilities of the ROSETok under different noise conditions. Overall, the contributions of this paper can be summarized as follows:

- We reveal the vulnerability of speech tokenizers and propose a transferable semantic attack method.
- We explore the effectiveness and transferability of our attack from the perspectives of token vulnerability, hierarchical representation drift, and LALM attention interference.
- Motivated by findings of the attack causes, we propose ROSETok, a speech tokenizer that improves noise-robust reconstruction and semantic preservation, enhancing the speech systems' robustness.
- We conduct extensive experiments to verify the efficiency, performance, and robustness of T-SemAttack. We verify the ROSETok's performance in reconstruction and downstream scenarios.

2 Background and Related Work

Speech Tokenizer and Speech Foundation Model. A speech tokenizer typically consists of an encoder and quantizer (decoder) [11]. The former encodes speech signals into representations, while

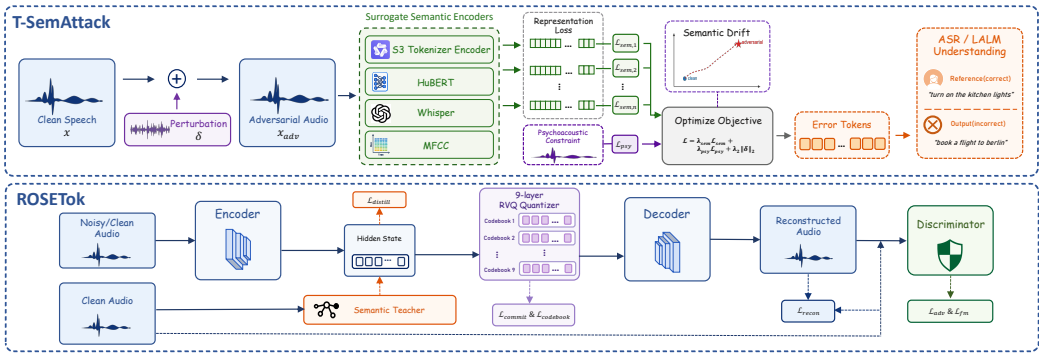


Figure 2: The architecture and pipeline of T-SemAttack and ROSETok.

the latter discretizes these representations into specific tokens, e.g., VQ- or FSQ-based tokens [12, 13]. As a speech front end, the speech tokenizer plays a crucial role in modern speech foundation models and has been widely adopted in various tasks and systems, including Qwen3-ASR [6] for ASR, CosyVoice [12] for TTS, and Qwen2-Audio [1] for LALMs. Since the speech tokenizer bridges the continuous acoustic space and the discrete semantic space, *safety vulnerability in this component may directly compromise the reliability and safety of the entire speech foundation model*.

Transferable ASR Attacks. Recent adversarial attacks on ASR systems have improved transferability and input-agnostic perturbation generation [14]. However, most existing methods attack the entire ASR pipeline, often relying on surrogate decoding objectives such as CTC loss [9], which can incur high optimization cost and obscure the role of front-end speech representations. Although recent work has observed instability in speech tokenizers [7], tokenizer-level transfer attacks and their propagation mechanisms remain underexplored. In contrast, we focus on the semantic encoder within the speech tokenizer and provide a systematic analysis of how small waveform perturbations induce semantic representation drift, token-level collapse, and downstream LALM attention disruption.

3 Transferable Semantic Attack

Most existing attacks [14, 15] against speech systems rely on specific output targets of ASR or LALM models and require access to the overall architecture, which limits efficiency and transferability. As the front end of LLM-based speech systems, the speech tokenizer plays a crucial role in providing modeling inputs for LLMs, making its security particularly important. We propose T-SemAttack, which attacks semantic representations by jointly perturbing surrogate semantic encoders and uses psychoacoustic constraints to enhance the perceptual quality in Figure 2.

3.1 Threat Model

Adversary Assumption. We consider a transfer-based attack scenario in which the adversary can access the input speech waveform and possesses a set of differentiable surrogate semantic encoders, but cannot access the internal gradients, parameters, or tokenizer details of the target systems. Under the constraint that the noise remains acceptable, the adversary aims to generate robust adversarial examples that cause speech systems to produce incorrect transcriptions. The adversarial examples should also be robust against adaptive defense, e.g., denoising and data transformations. Moreover, our attack is conducted in an offline setting with minimal real-time requirements.

Defense Assumption. We assume a strong defender who can detect the embedded noise and can use existing relevant tools or defense strategies to restore the original audio as much as possible.

3.2 Problem Definition

*Can perturbations disrupt the semantic representations of both continuous and discrete speech front ends **in a unified manner**, thereby affecting downstream speech models?*

Both continuous and discrete speech tokenizers typically produce continuous representations, with discrete tokenizers additionally applying quantization. Hard quantization is locally constant within quantization regions and non-differentiable at their boundaries. *We therefore target the continuous representations before any optional quantization*. Representation drift can alter the features supplied through a continuous interface or change discrete token assignments when the perturbed representations cross quantization boundaries. We evaluate the downstream effects empirically.

Given an original audio signal $x \in [-1, 1]^T$, where T denotes the audio length, our goal is to construct an adversarial example $x_{\text{adv}} = x + \delta$, where δ is a small adversarial perturbation. The overall optimization objective can be written as Eq. (1).

$$\begin{aligned} \min_{\delta} \quad & \mathcal{L} = \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}(x, x_{\text{adv}}) + \lambda_{\text{psy}}\mathcal{L}_{\text{psy}}(x, \delta) + \lambda_2\|\delta\|_2, \\ \text{s.t.} \quad & \|\delta\|_\infty \le \epsilon, \quad x_{\text{adv}} = x + \delta \in [-1, 1]^T, \end{aligned} \tag{1}$$

where $\mathcal{L}_{\text{sem}}$ measures the deviation of semantic representations, $\mathcal{L}_{\text{psy}}$ represents psychoacoustic masking, and $\|\delta\|_2$ controls perturbation level. We use $\lambda_{\text{sem}} = 100$, $\lambda_{\text{psy}} = 10^{-3}$, and $\lambda_2 = 0.1$.

3.3 Tokenizer-centric Semantic Attack

Representation Drifting. Representation drift refers to a shift in representations of perturbed speech relative to those of clean speech. T-SemAttack optimizes waveform perturbations to reduce the cosine similarity between these representations in the encoding space, causing different internal representations and thereby disrupting downstream models' understanding of speech content.

To improve cross-model transferability, we jointly optimize over multiple encoders. Specifically, we construct the following four surrogate branches: (1) **S3Tokenizer** [12], which directly simulates the semantic encoding process of the *discrete* speech tokenizer. (2) **HuBERT** [16], which provides deep semantic representations obtained through self-supervised learning. (3) **Whisper** encoder [17], which provides robust semantic representations learned from strong ASR pretraining. (4) Mel-Frequency Cepstral Coefficient (**MFCC**) [18, 9], is used as an additional low-level attack target [9, 8].

Let $\mathcal{F} = f_1, \ldots, f_K$ denote the K surrogate semantic encoders. For input x, encoder f_k outputs representations $f_k(x)$. Eq. (2) defines the semantic loss from frame-wise averaged cosine similarity.

$$\mathcal{L}_{\text{sem}} = \sum_{k=1}^{K} \text{CosSim}\left(f_k(x), f_k(x_{\text{adv}})\right). \tag{2}$$

By minimizing Eq. (2), the attack simultaneously drives shifts in multiple semantic spaces, thereby improving transferability to unseen ASR and LALM models.

Mechanistically, this attack can transfer to different ASR and LALM models because multiple semantic encoders, although trained with different objectives, share similar Mel-input bottlenecks and semantic structures. Therefore, disrupting the shared semantic representations in surrogate models often also disrupts the corresponding representations in unseen target models.

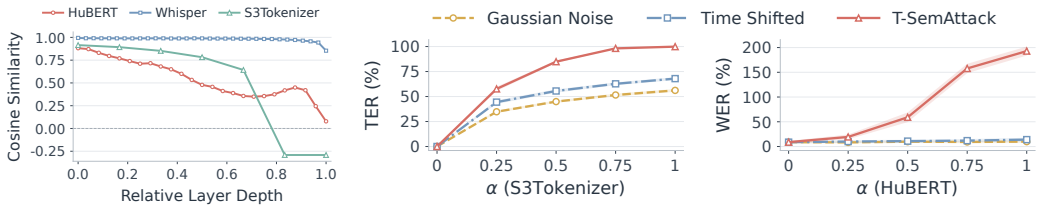
Psychoacoustic Model. Previous transferable ASR attacks suffer from low perceptual quality. We utilize a psychoacoustic constraint to ensure that the adversarial perturbation remains acceptable to humans. Given the original audio x, we first compute the corresponding masking threshold $\Theta(x)$ based on its short-time spectrum and use it as an upper-bound reference for the perturbation energy. Let $\text{PSD}(\delta)$ denote the power spectral density of the perturbation in the time-frequency domain. The psychoacoustic loss can then be written as Eq. (3).

$$\mathcal{L}_{\text{psy}} = \sum_{t,f} \max\left(0, \text{PSD}_{t,f}(\delta) - \Theta_{t,f}(x)\right), \tag{3}$$

which imposes a penalty only when the perturbation energy in a specific time-frequency unit exceeds the masking threshold. In this way, the optimization process preferentially hides the perturbation in frequency bands or local temporal regions where the original speech is stronger, thereby balancing attack effectiveness and perceptual quality. We follow prior work [19].

Optimization Process. In optimization, we extract multiple representations and update the adversarial audio based on the total loss in Eq. (1). The optimization adopts a PGD-style [20] projected update, where the perturbation is projected back onto the ℓ_∞ ball after each step in Eq. (4).

$$\delta \leftarrow \text{Proj}_{\|\delta\|_\infty \le \epsilon}\left(\delta - \alpha \cdot \text{sign}\left(\nabla_\delta \mathcal{L}\right)\right), \tag{4}$$



(a) Layer-wise similarity.

(b) Effect of perturbation dose controlled by α.

Figure 3: Layer-wise analysis of representation similarity and effectiveness of perturbation magnitude.

where α is the step size. In practice, we combine the SGD optimizer with an adaptive plateau learning-rate scheduler [9], which automatically reduces the learning rate when the loss stagnates, thereby improving optimization stability. The entire process is run for 500 iterations.

4 Why does the T-SemAttack work?

In Section 7, we conduct extensive experiments demonstrating the effectiveness of T-SemAttack. In this section, we explore why T-SemAttack is effective and how semantic attacks affect the model.

Layer-wise Representation Similarity. We compare the layer-wise representations of clean and attacked speech on LibriTTS. As shown in Figure 3a, HuBERT's cosine similarity decreases from 0.881 at the input embedding layer to 0.079 at the final layer. Whisper's shallow representations remain relatively stable, but its final-layer similarity still drops to 0.854. These results indicate that *small waveform perturbations can induce substantial shifts in deep representations*, highlighting the need for defenses that promote semantic stability throughout the encoding process.

Shared Attack Space. We extract a candidate shared subspace via singular value decomposition of normalized gradients and compare different attack directions under the same perturbation budget. Across 30 samples, using only the rank-1 shared direction yields a WER of 155.02% on the model HuBERT, comparable with the 191.64% achieved by the full ensemble attack and the 26.38% achieved by the residual attack. These findings suggest that shared directions can support strong cross-model transferability, while the contributions of different attack components depend on the target model.

Perturbation Magnitude. We test the attack direction and scale the perturbation according to $x_\alpha = x + \alpha\delta$, comparing it with energy-matched Gaussian noise and temporally shifted perturbations. As α increases from 0.5 to 0.75, HuBERT's WER rises from 59.05% to 157.74%, reaching 192.84% at $\alpha = 1$, compared with only 9.83% and 13.92% for the two controls, respectively. Meanwhile, at $\alpha = 1$, the token error rate (TER) of S3Tokenizer reaches 99.88%, exceeding the 56.12% observed for the Gaussian noise, respectively. These results indicate that *an appropriate direction can yield stronger attacks with less perturbation energy*.

5 Mechanism-Guided Robust Speech Tokenizer

In Section 4, we investigate why T-SemAttack can achieve effective attacks and reveal an important insight: *some existing speech tokenizers are not semantically robust and are vulnerable to adversarial attacks. We believe that defense solely at the LLM level is more time- and cost-intensive.* Therefore, we return to the core of the problem and aim to improve the stability of the tokenizer against noise. In this work, we utilize noisy training and semantic distillation to improve the robustness.

Semantic Preservation Module. Figure 3a illustrates the loss of deep semantic information. We employ a pretrained Semantic Teacher, Whisper [17], to supervise the representations of the model encoder. Specifically, we align the representations of noisy inputs with those of clean audio, thereby preserving semantic information and improving robustness to noise in downstream tasks.

Two-Stage Training. We utilize a two-stage training strategy consisting of base pretraining followed by attack-aware fine-tuning. In the first stage, we train on English and Chinese speech from Emilia, using speech corrupted by Gaussian noise as input and the corresponding clean audio as the reconstruction target, together with semantic distillation. We use a learning rate of 1×10^{-4} to

achieve high-fidelity reconstruction, semantic preservation, and general robustness to noise, yielding **ROSETok-Base**. In the second stage, to further adapt the model to structured semantic perturbations, we randomly select 2,000 utterances from LibriSpeech and generate perturbed inputs using T-SemAttack. We then perform Attack-Aware Fine-Tuning (AAF) on ROSETok-Base, using the original clean audio and its teacher representations as supervision. We reduce the learning rate to 5×10^{-6} to preserve the capabilities acquired during pretraining while improving robustness to adversarial perturbations, yielding **ROSETok-AAF**. Our model architecture is based on DAC and consists of an encoder, a nine-layer quantizer, and a decoder.

Training Objectives. We aim for high-fidelity reconstruction and semantic consistency. For both clean and noisy inputs, the model outputs high-quality clean audio. Moreover, we hope the encoder can preserve semantically rich information even for noisy audio inputs. We use a multi-frequency discriminator with hinge loss [21] for adversarial learning. The objective is in Eq. (5).

$$\mathcal{L}_{\text{tok}} = \lambda_{\text{distill}}\mathcal{L}_{\text{distill}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{cb}}\mathcal{L}_{\text{cb}} + \lambda_{\text{commit}}\mathcal{L}_{\text{commit}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{fm}}, \tag{5}$$

where $\mathcal{L}_{\text{distill}}$ is the semantic distillation loss, $\mathcal{L}_{\text{recon}}$ is the reconstruction loss, $\mathcal{L}_{\text{cb}}$ is the codebook loss, $\mathcal{L}_{\text{commit}}$ is the commitment loss, $\mathcal{L}_{\text{adv}}$ is the adversarial loss, and $\mathcal{L}_{\text{fm}}$ is the feature matching loss. Detailed descriptions are provided in Appendix E.3. The hyperparameters are set as $\{\lambda_{\text{distill}}, \lambda_{\text{recon}}, \lambda_{\text{cb}}, \lambda_{\text{commit}}\} = \{5.0, 15, 1.0, 0.25\}$. The weights of $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{fm}}$ are both 1.

6 Experimental Setup

6.1 Semantic Attack Setup

Datasets. To comprehensively evaluate the effectiveness of the semantic attack, we use bilingual and multi-speaker speech datasets. LibriTTS [22]: Following [9], we select single-speaker samples to evaluate the effectiveness of the attack in an English single-speaker scenario. Seed-TTS-Eval [23]: This is a widely used benchmark, containing both Chinese and English utterances from multiple speakers. We randomly sample 200 English utterances and 200 Chinese utterances for evaluation.

ASR Systems. We evaluate the attack performance on both open-source models (Qwen3-ASR [6], VibeVoice-ASR [24], FireRed-ASR [25], and Fun-ASR [26]) and commercial black-box models (iFlytek, Doubao ASR, and GPT-4o-mini-transcribe).

LALM Systems. The safety of LALMs is also a major focus of this work. We further evaluate our attack on several representative LALMs, including the open-source models Qwen2-Audio [1], Qwen3-Omni [2], and Audio-Flamingo 3 [5], as well as the black-box model Gemini-3-Flash.

Metrics. For efficiency, we report the Real-Time Factor (RTF) to measure the computational cost of the attack. For perceptual quality, we use Signal-to-Noise Ratio (SNR) and Mean Opinion Score (MOS) to evaluate the degree of distortion introduced by the attack in terms of audio quality. For attack effectiveness, we transcribe the adversarial audio using the target model and compute the Word Error Rate (WER). A higher WER indicates more severe corruption of the semantic content in the original speech.

6.2 Training Details of the Robust Tokenizer

Datasets. We train the ROSETok-Base on approximately 96.7k hours of bilingual speech data in Chinese and English. We utilize the English and Chinese subset from the Emilia [27] dataset.

Training Hyperparameters. We train the model on 8 NVIDIA H100 GPUs with a global batch size of 128. We use the AdamW optimizer, with an initial learning rate of 1×10^{-4} and $\beta_1 = 0.8, \beta_2 = 0.99$. The model is trained for a total of 200k steps, and the learning rate is decayed by 10%.

7 Experiments

Organization. First, we verify the effect and transferability of attacks on ASR systems. Second, we evaluate the impact of the attack on cross-modal audio understanding in LALMs. Third, we test the adaptive robustness of the attack against denoising, adversarial defense, and audio transformation techniques. For ROSETok, we evaluate its reconstruction stability on both clean and noisy audio. Finally, we conduct ablation studies for validation. Case study is provided in Appendix C.

Table 1: Comparison of attack efficiency, perceptual quality, and attack effectiveness across ASR systems and datasets. WER (%) is reported, where higher values indicate stronger attack effectiveness. The best results are highlighted and second-best results are underlined.

Methods	RTF(↓)	SNR(↑)	Surrogate Model		Open-source Model				Black-box Model			Avg.
			HuBERT	Whisper	Qwen3-ASR	VibeVoice	FireRed-ASR	Fun-ASR	iFlytek	Doubao	GPT-4o	
LibriTTS: A single-speaker English dataset.												
Clean	–	–	2.694	2.079	2.369	2.156	2.017	1.866	3.648	3.463	2.234	2.503
Gaussian Noise	0.003	15.197	3.228	2.213	2.583	2.736	2.933	2.488	3.837	3.691	6.677	3.376
E2E-VGuard	26.797	14.341	10.781	5.834	6.781	8.325	11.376	8.842	8.672	7.618	8.030	8.473
AudioShield	–	-2.440	16.007	10.363	15.450	15.241	13.800	12.171	16.815	14.273	11.049	13.908
T-SemAttack	21.800	**16.319**	**156.913**	**111.367**	**27.388**	**63.555**	**33.529**	**43.679**	**49.607**	**20.004**	**46.089**	**61.348**
Seed-TTS (EN): A multi-speaker English dataset.												
Clean	–	–	6.286	0.851	1.040	1.433	5.906	1.015	1.942	3.001	1.907	2.598
Gaussian Noise	0.003	15.228	10.739	1.307	1.695	3.862	10.985	2.061	3.457	2.679	9.849	5.182
E2E-VGuard	26.797	12.907	42.084	15.913	15.643	26.609	35.936	20.541	19.051	14.233	18.374	23.154
AudioShield	–	-2.364	62.779	42.840	**50.674**	51.296	54.301	56.806	54.181	**61.297**	42.172	52.927
T-SemAttack	21.800	**17.598**	**189.289**	**125.913**	37.445	**103.887**	**65.406**	**63.779**	**71.726**	29.948	**49.457**	**81.872**
Seed-TTS (ZH): A multi-speaker Chinese dataset.												
Clean	–	–	–	3.086	0.790	1.521	1.760	0.614	1.560	1.109	2.984	1.678
Gaussian Noise	0.003	15.387	–	5.434	1.207	6.171	2.899	1.466	1.331	0.774	15.129	4.301
E2E-VGuard	26.797	13.607	–	18.392	5.322	27.258	8.896	6.879	3.119	1.944	18.279	11.261
AudioShield	–	-2.435	–	28.113	**19.998**	50.729	16.609	23.267	14.345	**30.221**	37.687	27.621
T-SemAttack	21.800	**17.652**	–	**109.100**	18.833	**144.637**	**19.293**	**29.021**	**19.214**	6.298	**60.116**	**50.814**

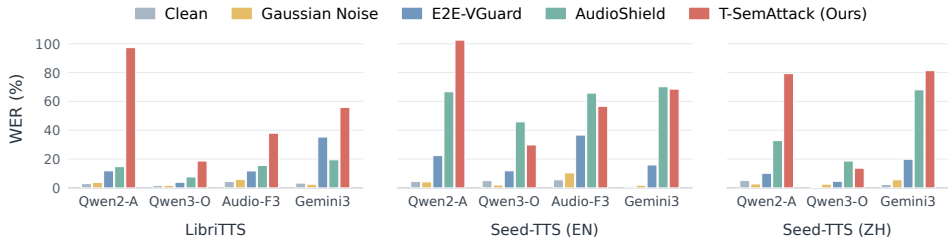


Figure 4: ASR performance across different datasets and LALMs, measured by WER (%).

7.1 Evaluation on ASR Systems

In Table 1, T-SemAttack achieves a better balance among efficiency, perceptual quality, and attack strength. Its RTF is 21.800, lower than E2E-VGuard's 26.797, while maintaining an SNR of 16.319 with good perceptual quality. Subjective perceptual experiments provided in Appendix D show T-SemAttack achieve human acceptable perception. Details and results of AudioShield's efficiency comparison are provided in Appendices A and E.5. AudioShield can produce relatively high WERs, but it almost destroys the audio quality, making the content difficult to hear clearly (Table 8). In contrast, with better perceptual quality, T-SemAttack increases the WERs of multiple open-source and black-box ASR systems from about 1%∼4% on clean speech to 20.004%∼63.555%, reaching 49.607% and 46.089% on iFlytek and GPT-4o-mini-transcribe, respectively.

Table 1 verifies generalization in multi-speaker and bilingual scenarios. On Seed-TTS (EN), T-SemAttack achieves an average WER of 81.872%, higher than E2E-VGuard's 23.154%. On the Chinese subset, the average WER remains 50.814%. Notably, T-SemAttack achieves a cross-linguistic attack on the Chinese dataset because our surrogate model, (Whisper and S3Tokenizer) are trained on large-scale multilingual datasets. Commercial systems also show robustness differences in Chinese scenarios. This experiment indicates that the transferability of semantic attacks generally holds.

7.2 Evaluation on LALMs

For evaluation on LALMs, we consider two tasks, i.e., ASR and question-answering tasks.

ASR Evaluation. Figure 4 shows the impact of T-SemAttack on the speech understanding ability of LALMs. Compared with ASR systems, LALMs not only need to recognize speech content, but also *align audio tokens with the text reasoning process*. Therefore, they are more sensitive to semantic drift at the tokenizer level. T-SemAttack achieves an average WER of 58.30% over 11 dataset-model combinations, which is much higher than E2E-VGuard's 16.76% and AudioShield's 38.66%. For example, on LibriTTS, the WER of Qwen2-Audio increases from 2.982% to 97.280%. On the Chinese set, Qwen2-Audio and Gemini3-Flash also reach 79.259% and 81.300%, respectively. These

Table 2: **Robustness evaluation** against advanced denoising models across two models, adversarial audio defense, and audio transformations on the VibeVoice-ASR with averaged WER(%) values.

Method	Denoiser		DeepFilterNet3		Adversarial Defense			Diffusion	Audio Transformation		
	Whisper	VibeVoice	Whisper	VibeVoice	Resample	Mel	Q-D	AudioPure	Speed	PS	Band-filter
Gaussian Noise	2.964	1.720	2.468	1.133	5.256	6.569	2.984	8.737	57.517	3.735	20.334
E2E-VGuard	10.205	7.276	10.168	7.148	13.841	17.650	9.447	13.898	67.112	10.282	28.863
AudioShield	9.930	9.738	10.741	9.187	16.030	16.751	14.673	**37.497**	61.546	13.379	23.901
T-SemAttack	**96.763**	**42.901**	**113.997**	**37.805**	**55.805**	**58.997**	**73.364**	21.571	**89.898**	**70.635**	**71.085**

results show that once the semantic tokenizer produces incorrect semantic tokens, subsequent LALMs can hardly recover the original semantics through stronger language modeling.

This phenomenon is also consistent with the attention analysis in Section F.2 which shows that attacked audio tokens weaken the model's reliance on audio context, causing cross-modal attention to shift more toward the text itself and thereby amplifying semantic understanding errors.

QA Evaluation. we evaluate four LALMs on the *audio question-answering (QA)* task on the AIR-Bench [28] with randomly selected 100 samples. We use DeepSeek-V4-Flash as the judge model to assign a Response Quality Score (RQS), where higher scores indicate better response quality. As shown in Table 3, all four LALMs receive lower RQS values under T-SemAttack, with the mean across models decreasing from 7.58 on clean inputs to 5.48 on perturbed inputs. These results show that T-SemAttack can also degrade the quality of LALMs' audio QA performance.

Table 3: Response Quality Scores of LALMs on AIR-Bench with and without T-SemAttack. Avg. denotes the mean across the four models.

Input	Qwen2-Audio	Qwen3-Omni	Audio-Flamingo 3	Gemini 3 Flash	Avg. (↓)
Clean	6.63	7.73	8.14	7.81	7.58
T-SemAttack	**4.58**	**5.23**	**6.20**	**5.89**	**5.48**

7.3 Adaptive Defense against Semantic Attack

We consider some strong defense settings: denoising, adversarial defense, diffusion-based adversarial defense, and audio transformations. This experiment is conducted on the Seed-TTS-EN dataset.

Advanced Denoising. We utilize two SOTA denoising models, Denoiser [29] and DeepFilterNet3 [30]. The results in Table 2 show that after Denoiser processing, T-SemAttack still achieves WERs of 96.763% and 42.901% on Whisper, and VibeVoice-ASR, respectively. The WERs remain high after DeepFilterNet3 processing. In contrast, under the same defenses, the WERs of E2E-VGuard and AudioShield are mostly only 7%$\sim$16%.

Adversarial Defense and Audio Transformations. We consider adversarial defense based on WaveGuard, diffusion-based defense (AudioPure [31]), and audio transformation-based defense. The results are shown in Table 2. The best defense method is AudioPure, achieving a WER score of 21.571%, and still misleads correct transcription. The highest WER of AudioPure-perturbed samples is 91.667%, reflecting that this method cannot perfectly disrupt the T-SemAttack. We find that previous methods are not robust enough against these advanced defensive methods, and T-SemAttack achieves SOTA performance. The explanations of these techniques are provided in the Appendix E.2.

These results suggest that T-SemAttack perturbations are not merely removable noise by conventional speech enhancement. Although Denoiser and DeepFilterNet3 can suppress the acoustic disturbance, they cannot reliably recover the original token sequence or semantic representation once the tokenizer output has been shifted. This indicates that speech enhancement alone is insufficient, and *ROSETok must explicitly model both noise robustness and semantic consistency at the tokenizer level.*

7.4 Robust Speech Tokenizer Analysis

Reconstruction Performance. We first evaluate speech tokenizers' reconstruction performance on Seed-TTS (EN) using Short-Time Objective Intelligibility (STOI). As shown in Table 4, ROSETok-Base achieves reconstruction scores of 0.966, 0.950, and 0.842 on clean, Gaussian-noise-corrupted,

Table 4: Reconstruction and downstream evaluation of speech tokenizers under different inputs.

Model	N_q	RTF(↓)	Reconstruction(↑)			Downstream SpeechLLM(↑)		
			Clean	Gaussian	T-SemAttack	Clean	Gaussian	T-SemAttack
Acoustic and Semantic Speech Tokenizer								
SNAC	4	0.003	0.874	0.817	0.782	28.3	5.6	14.6
XY-Tokenizer	8	0.025	0.905	0.849	0.762	92.3	87.8	23.1
X-Codec 2	1	0.021	0.864	0.808	0.771	20.1	8.1	7.4
DAC	9	0.008	0.943	0.891	0.830	25.1	9.1	13.9
Robust Speech Tokenizer								
MOSS-Audio-Tokenizer	9	0.029	0.935	0.894	0.831	54.1	27.6	27.4
StableToken (Pre-Q)	1	0.175	0.647	0.647	0.592	**96.7**	**95.4**	35.0
StableToken (Post-Q)						91.0	89.5	27.9
ROSETok-Base	9	**0.010**	**0.966**	**0.950**	0.842	93.0	92.9	62.4
ROSETok-AAF		–	0.945	0.937	**0.903**	92.7	89.3	**77.9**

Note. Pre-Q and Post-Q denote representations before and after quantization for downstream evaluation. All other models use Pre-Q features. Bold metric values indicate the best reported result in each column.

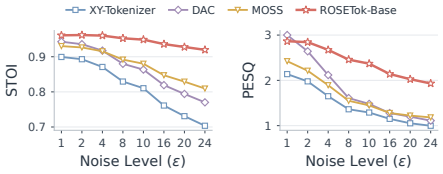


Figure 5: Noise-level evaluation.

Table 5: Ablation results on reconstruction and semantic evaluation of the FSC task with MLP. "w/o" represents without this loss/component.

Method	Recon-Clean		Recon-Noisy		Semantic-FSC(↑)	
	STOI(↑)	PESQ(↑)	STOI(↑)	PESQ(↑)	Clean	Noisy
w/o noisy	**0.956**	**2.892**	0.879	1.578	**69.259**	11.363
w/o $\mathcal{L}_{\text{distill}}$	0.947	2.652	**0.934**	**2.322**	7.118	5.220
Ours	0.945	2.573	0.931	2.304	61.139	**60.032**

and T-SemAttack-perturbed inputs, respectively, outperforming baselines. ROSETok-AAF further improves the reconstruction score under T-SemAttack to 0.903 while maintaining high reconstruction quality on clean and Gaussian-noise-corrupted inputs. These results demonstrate that two-stage training can preserve reconstruction quality while improving adversarial robustness.

Downstream Semantic Robustness. Following the XARES-LLM benchmark [32], we feed each tokenizer's pre-quantized continuous representations into SmolLM2-135M and evaluate their impact on downstream semantic understanding using the FSC task. We train the LLM with clean samples and test with different inputs. Existing tokenizers generally exhibit substantial degradation under semantic attacks. For example, StableToken achieves scores of 96.7 and 95.4 on clean and Gaussian-noise-corrupted inputs, respectively, but only 35.0 under T-SemAttack, indicating that robustness to Gaussian noise is insufficient to ensure downstream semantic understanding. In comparison, ROSETok-Base achieves 62.4 under attack, and ROSETok-AAF further improves this score to 77.9. *These results show that the robustness of front-end representations affects downstream performance and that AAF can mitigate performance degradation from semantic attacks.*

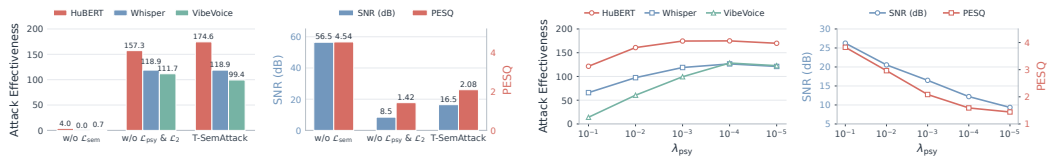
Various Noise-Level Reconstruction. We evaluate models' noise robustness under Gaussian noise with magnitudes from $1/255$ to $24/255$ as shown in Figure 5. Most models suffer severe performance degradation under stronger noise, e.g., DAC drops from 0.943 on clean input to 0.770 at the $24/255$ noise level. In contrast, ROSETok-Base remains relatively stable and still achieves good reconstruction quality under strong noise (24/255), with a 0.920 STOI score.

7.5 Ablation Study

In this section, we conduct ablation experiments. For the attack, we randomly select 30 samples from Seed-TTS (EN). For ROSETok, we train on the LibriSpeech [33] for 100k steps.

Component Effect of T-SemAttack. We study the effect of each component in Eq. (1) of T-SemAttack. Results are shown in Figures 6a and 6b. We find that, after removing the semantic attack loss $\mathcal{L}_{\text{sem}}$, the perceptual quality remains high, but the attack is almost ineffective on HuBERT, Whisper, and VibeVoice-ASR, with an attack success rate of only about 1.556%. After removing the perceptual loss $\mathcal{L}_{\text{psy}}$ and $\mathcal{L}_2$, the SNR drops to 8.5, and the PESQ score falls below 2.08. Combining both components achieves a sound balance between perceptual quality and attack effectiveness.

Perception and Attack Balance. In Eq. (1), λ_{psy} balances attack effectiveness and perceptual quality. We evaluate its sensitivity over the range from 10^{-1} to 10^{-5}. Results in Figure 6c6d show



(a) Attack (b) Perception (c) λ_{psy} (Attack) (d) λ_{psy} (Perception)

Figure 6: Ablation study for the T-SemAttack. (a)(b) Component Effect. (c)(d) Sensitivity to λ_{psy}.

that when λ_{psy} is set to 10^{-3} or smaller, the attack effectiveness on the three models shows no significant improvement, while the SNR continues to decrease. Therefore, we set the λ_{psy} to 10^{-3}.

Component and Semantic Effect of ROSETok. We verify the effect of ROSETok's components at noise levels of $2/255 \sim 16/255$. As shown in Table 5, removing noisy training reduces noisy-input STOI from 0.931 to 0.879 and PESQ from 2.304 to 1.578, confirming the importance of noisy-clean training. To evaluate semantic capability and consistency, we adapt the ROSETok encoder to the XARES [34] benchmark. Higher benchmark scores are better. On Fluent Speech Commands (FSC), a small-scale ASR task, we train on the clean training set and evaluate on its clean and noisy test sets to assess the model's robustness to noise in downstream scenarios. When the noisy signal input is removed, the FSC score drops from 60.032 to 11.363. Removing semantic supervision leads to even more severe semantic degradation on FSC. Overall, ROSETok achieves high-fidelity reconstruction of both clean and noisy audio while preserving strong semantic robustness under noisy conditions.

Effect of the Semantic Teacher. Semantic teacher encoder is used in the training stage of ROSETok. We test the effect of the semantic teacher in ROSETok-Base by replacing Whisper with HuBERT while keeping the remaining training configuration unchanged. Table 6 reports reconstruction and denoising performance, together with downstream semantic performance on Fluent Speech Commands (FSC). The two teachers yield comparable reconstruction and denoising scores on clean and noisy inputs. However, the downstream FSC score with XARES [34] probe decreases from 73.90 with Whisper to 31.48 with HuBERT. These results support the choice of Whisper as the semantic teacher in our setting, despite the similar reconstruction quality obtained with either teacher.

Table 6: Effect of the semantic teacher on ROSETok-Base reconstruction and downstream semantic performance.

Teacher	Clean inputs		Noisy inputs		FSC (MLP)
	STOI(↑)	PESQ(↑)	STOI(↑)	PESQ(↑)	FSC(↑)
Whisper	0.965	3.063	0.950	2.431	**73.90**
HuBERT	0.965	3.088	0.950	2.453	31.48

8 Conclusion

In this work, we explore the robustness of speech tokenizers. Based on the semantic encoder, we propose T-SemAttack, which effectively attacks current mainstream ASR and LALM systems. We further analyze the causes of the attack and, inspired by these findings, propose a robust speech tokenizer that enables high-quality reconstruction from noisy speech. These findings highlight the speech tokenizer as a critical safety bottleneck in modern speech foundation models. We hope our analysis and ROSETok can provide useful insights for developing future speech tokenizers with stronger semantic consistency and robustness.

Acknowledgments and Disclosures

We sincerely thank the reviewers for their valuable suggestions. This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (JCYJ20220818101014030), and Fundamental Research Funds for the Central Universities under Grant No. 2024ZCJH05. The authors declare no competing interests relevant to this work.

References

[1] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[2] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[3] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International conference on machine learning*, pages 5530–5540. PMLR, 2021.

[4] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[5] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

[6] Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

[7] Yuhan Song, Linhao Zhang, Chuhan Wu, Aiwei Liu, Wei Jia, Houfeng Wang, and Xiao Zhou. Stabletoken: A noise-robust semantic speech tokenizer for resilient speechllms. *arXiv preprint arXiv:2509.22220*, 2025.

[8] Zheng Fang, Tao Wang, Lingchen Zhao, Shenyi Zhang, Bowen Li, Yunjie Ge, Qi Li, Chao Shen, and Qian Wang. Zero-query adversarial attack on black-box automatic speech recognition systems. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 630–644, 2024.

[9] Zhisheng Zhang, Derui Wang, Yifan Mi, Zhiyong Wu, Jie Gao, Yuxin Cao, Kai Ye, Minhui Xue, and Jie Hao. E2e-vguard: Adversarial prevention for production llm-based end-to-end speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[10] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

[11] Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, Haibin Wu, Haici Yang, Anastasia Kuznetsova, Artem Ploujnikov, Ricard Marxer, et al. Discrete audio tokens: More than a survey! *arXiv preprint arXiv:2506.10274*, 2025.

[12] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

[13] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[14] Weifei Jin, Yuxin Cao, Junjie Su, Derui Wang, Yedi Zhang, Minhui Xue, Jie Hao, Jin Song Dong, and Yixian Yang. Whispering under the eaves: protecting user privacy against commercial and llm-powered automatic speech recognition systems. In *Proceedings of the 34th USENIX Conference on Security Symposium*, pages 4643–4662, 2025.

[15] Guangke Chen, Fu Song, Zhe Zhao, Xiaojun Jia, Yang Liu, Yanchen Qiao, Weizhe Zhang, Weiping Tu, Yuhong Yang, and Bo Du. Audiojailbreak: Jailbreak attacks against end-to-end large audio-language models. *IEEE Transactions on Dependable and Secure Computing*, 2026.

[16] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[18] Yunjie Ge, Lingchen Zhao, Qian Wang, Yiheng Duan, and Minxin Du. Advddos: Zero-query adversarial attacks against commercial speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 18:3647–3661, 2023.

[19] Meng Chen, Li Lu, Junhao Wang, Jiadi Yu, Yingying Chen, Zhibo Wang, Zhongjie Ba, Feng Lin, and Kui Ren. Voicecloak: Adversarial example enabled voice de-identification with balanced privacy and utility. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(2):1–21, 2023.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[21] Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, et al. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025.

[22] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

[23] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

[24] Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, et al. Vibevoice-asr technical report. *arXiv preprint arXiv:2601.18184*, 2026.

[25] Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*, 2025.

[26] Keyu An, Yanni Chen, Zhigao Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Bo Gong, Xiangang Li, Yabin Li, et al. Fun-asr technical report. *arXiv preprint arXiv:2509.12508*, 2025.

[27] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE, 2024.

[28] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, 2024.

[29] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real time speech enhancement in the waveform domain. In *Interspeech*, 2020.

[30] Hendrik Schröter, Tobias Rosenkranz, Alberto N. Escalante-B., and Andreas Maier. DeepFilterNet: Perceptually motivated real-time speech enhancement. In *INTERSPEECH*, 2023.

[31] Shutong Wu, Jiongxiao Wang, Wei Ping, Weili Nie, and Chaowei Xiao. Defending against adversarial audio via diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[32] Heinrich Dinkel, Jiahao Zhou, Guanbo Wang, Yadong Niu, Junbo Zhang, Yufeng Hao, Ying Liu, Ke Li, Wenwu Wang, Zhiyong Wu, et al. The interspeech 2026 audio encoder capability challenge for large audio language models. *arXiv preprint arXiv:2603.22728*, 2026.

[33] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[34] Junbo Zhang, Heinrich Dinkel, Yadong Niu, Chenyu Liu, Si Cheng, Anbei Zhao, and Jian Luan. X-ares: A comprehensive framework for assessing audio encoder performance. In *Interspeech 2025*, 2025.

[35] Zhisheng Zhang, Derui Wang, Qianyi Yang, Pengyang Huang, Junhan Pu, Yuxin Cao, Kai Ye, Jie Hao, and Yixian Yang. {SafeSpeech}: Robust and universal voice protection against malicious speech synthesis. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 4581–4600, 2025.

[36] Jiaqi Li, Yao Qian, Yuxuan Hu, Leying Zhang, Xiaofei Wang, Heng Lu, Manthan Thakker, Jinyu Li, Sheng Zhao, and Zhizheng Wu. Flexicodec: A dynamic neural audio codec for low frame rates. In *The Fourteenth International Conference on Learning Representations*, 2026.

[37] Heinrich Dinkel, Xingwei Sun, Gang Li, Jiahao Mei, Yadong Niu, Jizhong Liu, Xiyang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, et al. Dashengtokenizer: One layer is enough for unified audio understanding and generation. *arXiv preprint arXiv:2602.23765*, 2026.

[38] Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*, 2025.

[39] Shehzeen Hussain, Paarth Neekhara, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. {WaveGuard}: Understanding and mitigating audio adversarial examples. In *30th USENIX security symposium (USENIX Security 21)*, pages 2273–2290, 2021.

[40] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

[41] Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer. Snac: Multi-scale neural audio codec. *arXiv preprint arXiv:2410.14411*, 2024.

[42] Yitian Gong, Luozhijie Jin, Ruifan Deng, Dong Zhang, Xin Zhang, Qinyuan Cheng, Zhaoye Fei, Shimin Li, and Xipeng Qiu. Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs. *arXiv preprint arXiv:2506.23325*, 2025.

[43] Yitian Gong, Kuangwei Chen, Zhaoye Fei, Xiaogui Yang, Ke Chen, Yang Wang, Kexin Huang, Mingshu Chen, Ruixiao Li, Qingyuan Cheng, et al. Moss-audio-tokenizer: Scaling audio tokenizers for future audio foundation models. *arXiv preprint arXiv:2602.10934*, 2026.

A Discussion and Limitations

In this section, we discuss some important aspects.

Q1: Can T-SemAttack's perturbation be really imperceptible to humans?

A: In the setting considered in this paper, what we aim to emphasize is that the noise does not compromise the normal usability of the original audio, that is, its **harmlessness**, rather than true imperceptibility. T-SemAttack adopts a linear additive strategy and incorporates the masking effect to minimize its impact on the original audio content as much as possible. Subjective experiments in Appendix D show that T-SemAttack achieves better audio quality and content consistency than baselines, achieving acceptable and normal use of the attacked audio with a MOS-Q of 3.01. Prior research [35] suggests that MOS scores above 3 indicate noise levels acceptable to human listeners. T-SemAttack's noise level is comparable to that of **a typical cafe or street environment** where people routinely listen to speech. In addition, T-SemAttack exhibits strong robustness, which ensures that even if the defender perceives the embedded noise, denoising methods, adversarial defenses, and data transformations cannot effectively remove or neutralize it. Even when adversarial audio is detected, defenders lack effective mitigation strategies. Moreover, because the audio must be retained, e.g., recordings in legal contexts, defenders cannot remove the perturbations by replacing it with a new recording. *Therefore, T-SemAttack aims to avoid disrupting the normal use of audio, rather than to achieve true imperceptibility*.

Q2: How efficient is T-SemAttack?

A: First, we focus on an offline scenario, where the real-time requirement is relatively low. In Table 1, we evaluate the RTF of our method and the baselines. Specifically, the RTF is calculated as the total time required by each method to attack the same LibriTTS dataset divided by the total duration of the audio. Details of efficiency computation are provided in Appendix E.5.

T-SemAttack achieves an RTF of 21.8, outperforming E2E-VGuard of 26.797. This is because our attack targets the semantic encoder rather than the entire model, which partially improves efficiency. It is worth noting that AudioShield is a real-time attack framework, but it requires a very long initialization stage. Specifically, before AudioShield can be used to attack audio, it must first be trained on LibriSpeech for an extended period of time. Therefore, the RTF of AudioShield is 87.045 for warmup-based training and 0.040 for each sample attack after long-time offline training.

Therefore, although T-SemAttack currently achieves better time efficiency than E2E-VGuard, improving its real-time performance remains a direction for future work, e.g., multi-GPU parallelization.

Q3: Why does ROSETok preserve semantics at the encoder-level rather than the token-level?

A: Token-level stability and encoder-level semantic preservation address different failure modes. Token consistency encourages discrete-code invariance, while ROSETok targets the semantic representation drift identified by T-SemAttack. Since token changes do not always imply semantic degradation, we align encoder representations with clean semantic teachers rather than enforcing exact token invariance. Specifically, we consider two reasons.

First, this approach has broader applicability. Some recent codecs [36] introduce semantic distillation in the first few layers of RVQ and use the subsequent layers to model acoustic information. Other methods [7] directly constrain the stability of quantized tokens under noisy inputs. In contrast, ROSETok directly constrains the semantic representations output by the encoder, encouraging them to remain close to the semantic representations of clean speech under noisy inputs. This not only provides more robust inputs for the subsequent quantizer, but also makes the approach applicable to recent speech tokenizers [37] that are based on continuous representations rather than discrete tokens. Accordingly, ROSETok is designed precisely to align the semantic representations of noisy inputs with those of clean audio using a Whisper teacher.

Second, changes in tokens do not necessarily imply changes in semantics. Random Gaussian noise may cause a large proportion of the tokenizer's output tokens to change. However, if the deep semantic representations remain stable, downstream ASR or LALM models may still recover the correct semantics through their own modeling capabilities. By contrast, the key mechanism of T-SemAttack is not simply to change token IDs, but to drive semantic drift in the encoder representations, making it difficult for downstream models to recover the original content.

Therefore, the goal of ROSETok is not to keep tokens unchanged, but to align the semantic representations of perturbed inputs with those of clean speech. Compared with token-level consistency, encoder-level semantic preservation corresponds more directly to the attack mechanism identified in this paper and, in this work, serves as a more suitable training objective for robust speech tokenizers.

B Ethical Considerations and Broader Impacts

Dual-Use and Disclaimer of the T-SemAttack. We recognize that T-SemAttack is a dual-use technique. While this work exposes security threats in existing speech tokenizers and encourages researchers to develop more robust speech tokenizers, the same technique could be misused to degrade automatic transcription, evade speech-based moderation or auditing systems, interfere with voice assistants, or reduce the reliability of speech interfaces used in accessibility, documentation, or security-sensitive applications. We therefore treat the proposed attack as a security research artifact rather than a deployable tool. To mitigate unforeseen impacts, before releasing the source code, we will publish a disclaimer stating that our tool is intended solely for academic research and related purposes. Users must not use this tool for unauthorized or illegal activities, and all actions taken by users are unrelated to the creators and publishers of the tool.

Use of Third-Party Systems. Our black-box evaluation uses commercial ASR and LALM APIs only to assess transferability. We do not attempt to bypass authentication, rate limits, monitoring, or any access-control mechanism. The evaluation is conducted at a limited scale, uses only public benchmark utterances, and does not involve private user speech. We report aggregate performance metrics and do not use the services to extract model parameters, training data, or proprietary information.

Subjective Experiments. We obtain informed consent from all participants during recruitment, and all selected participants are over 18 years old. The identities of the participants are anonymized. We do not collect any information beyond what is necessary for the subjective experiments, and the relevant information and responses are used only for academic research. We will not disclose any such sensitive information. All adversarial audio samples are discarded after the subjective experiments are completed, so they will not cause any additional commercial or societal impact.

Defensive Intent and Limitations. We also release and evaluate ROSETok to encourage defensive research on robust speech tokenizers. Nevertheless, ROSETok should not be interpreted as a complete mitigation against all semantic attacks. Its robustness depends on the training distribution and threat model, and deployed systems should combine tokenizer-level robustness with additional monitoring, denoising, adversarial evaluation, and system-level safeguards.

C Case Study: T-SemAttack in Text-To-Speech for Protection

To evaluate the potential application of T-SemAttack in speech protection scenarios, we conduct a case study on text-to-speech (TTS) systems. Specifically, we consider a realistic threat in which an attacker can obtain a user's speech samples and utilize advanced TTS systems for voice cloning [9]. If the original speech is slightly perturbed before being released or uploaded in a form that remains acceptable to humans, downstream speech generation systems would struggle to accurately extract speaker characteristics or semantic information. Motivated by this scenario, we evaluate whether T-SemAttack can serve as a speech protection against the unauthorized use of TTS systems.

We select three representative TTS systems for evaluation, i.e., CosyVoice [12], CosyVoice2 [13], and Step-Audio-TTS [38]. We evaluate the generated speech from two perspectives: WER and SIM. A higher WER indicates more severe semantic deviation in the generated content, while a lower SIM indicates that the system has greater difficulty preserving the voice characteristics of the original speaker. We conduct experiments on the Seed-TTS (EN) dataset, using the ground-truth transcription of each audio sample as the input reference text.

Table 7 presents the results of this case study. With clean input, all three TTS systems preserve the semantic content well, achieving WERs of only 4.583, 4.484, and 3.734, respectively. These observations suggest that these systems can effectively use the speaker features contained in the input speech for speech generation. In contrast, when the input speech is processed by T-SemAttack, the generation quality of all TTS systems degrades substantially. Step-Audio-TTS is affected most severely, with its WER increasing sharply from 3.734 to 122.746. On average, the WER increases

Table 7: Case study of T-SemAttack in the TTS protection scenario. The higher WER(%) represents more severe semantic errors in the generated speech. The lower the SIM, the lower the similarity between the generated speech and the original speaker.

Method	CosyVoice [12]		CosyVoice2 [13]		Step-Audio-TTS [38]	
	WER(↑)	SIM(↓)	WER(↑)	SIM(↓)	WER(↑)	SIM(↓)
Clean	4.583	0.599	4.484	0.836	3.734	0.772
T-SemAttack	**52.476**	**0.212**	**33.427**	**0.330**	**122.746**	**0.299**

from 4.267 under the clean condition to 69.550, while the speaker similarity decreases from 0.736 to 0.280. These results show that T-SemAttack can disrupt the modeling of speech content in TTS.

This result suggests that semantic-level perturbations can also affect the conditional modeling process in TTS systems. When a TTS system relies on input speech to extract semantic or speaker representations, the semantic representation disruption and tokenizer-level representation perturbation caused by T-SemAttack can affect the generation stage, resulting in content errors and reduced speaker similarity in the generated speech. Therefore, this case study provides another perspective supporting the critical role of speech tokenizers in modern speech systems.

D Human Study

To evaluate the human perception of the perturbation, we conduct a subjective human listening study to assess the perceptual quality of the attacked audio generated by T-SemAttack and the baselines. This section details the recruitment process, design of the questionnaire, and results of evaluations.

Recruitment Process. This subjective survey has been approved by the Human Ethics Research Committee at the first author's institution. We create the questionnaire through the Credamo platform and recruit 20 volunteers to participate in the survey, which is comparable to similar studies, e.g., AntiFake (24 participants) and E2E-VGuard (36 participants). All volunteers are over 18 years old and are English-educated. Their average response time is ∼330 seconds, providing reliable results.

Questionnaire Setup. We randomly select four speakers (two male and two female) from the Seed-TTS-EN dataset, with one utterance per speaker. Each utterance is processed by baselines, i.e., E2E-VGuard and AudioShield , as well as our T-SemAttack, resulting in 12 attacked audio samples and 4 original samples. For each attacked sample, participants are asked to compare it with the corresponding original audio and provide ratings from three perspectives: (1) **Audio Quality**, which measures the noise level and overall perceptual quality of the attacked audio; (2) **Content Consistency**, which evaluates whether the linguistic content of the attacked audio remains aligned with the original; and (3) **Timbre Similarity**, which assesses whether the speaker identity in the attacked audio is preserved relative to the original.

All three aspects are scored on a 0∼5 point Mean Opinion Score (MOS) scale, where higher scores indicate better perceptual quality. To improve the confidence of the subjective evaluation and reduce bias, we report the MOS values together with their 95% confidence intervals [35].

Table 8: Subjective evaluation from audio quality, content intelligibility, and speaker similarity aspects with 95% confidence intervals.

Method	**MOS-Q**(↑)	**MOS-C**(↑)	**MOS-S**(↑)
E2E-VGuard	2.67±0.21	4.29±0.17	**3.98±0.20**
AudioShield	2.63±0.30	3.32±0.31	1.66±0.33
T-SemAttack	**3.01±0.28**	**4.32±0.19**	3.87±0.25

Results and Analysis. The detailed MOS results and analysis are provided in Table 8. We evaluate three dimensions: overall quality (MOS-Q), content intelligibility (MOS-C), and speaker similarity (MOS-S). Across all three dimensions, our method, T-SemAttack, consistently achieves the highest or comparable MOS values among all methods. In particular, AudioShield introduces a clearly perceivable perturbation that almost prevents listeners from recognizing the linguistic content of the

original audio, as reflected by its low MOS-Q (2.63), MOS-C (3.32), and MOS-S (1.66). We attribute this to the fact that AudioShield is trained on LibriSpeech and has fitted a substantial amount of female voice characteristics, and that modifying the latent representation cannot guarantee the fidelity of the reconstructed waveform. E2E-VGuard better preserves content intelligibility with a MOS-C of 4.29 and a MOS-S of 3.98 for speaker similarity, but still introduces audible artifacts that lower its overall quality with a MOS-Q of 2.67. In contrast, T-SemAttack adopts a linear additive strategy combined with the psychoacoustic masking effect, which minimizes its impact on the original audio content. It achieves the highest MOS-Q (3.01), *which surpasses the 3.0 threshold generally regarded as acceptable perceptual distortion* [35], and maintains strong MOS-C (4.32) and MOS-S (3.87). These subjective experiments show that T-SemAttack does not compromise the linguistic content or speaker timbre, thereby preserving the normal usability of the audio.

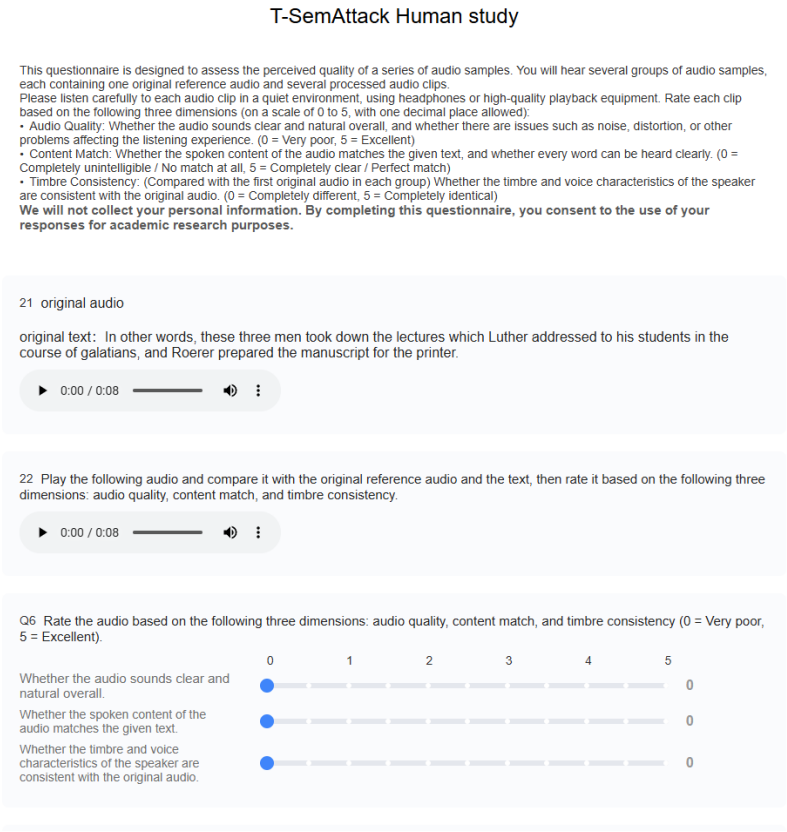


Figure 7: The screenshot of the questionnaire for subjective experiments.

The screenshot and full text of the questionnaire are shown in Figure 7.

E Additional Details

E.1 Attack Details

Table 9 summarizes the main hyperparameters used throughout T-SemAttack. "lr" refers to the initial learning rate of the SGD optimizer, used to control the update step size of the adversarial noise during backpropagation. "lr_min" refers to the minimal learning rate. "max_steps" denotes the total optimization steps for the T-SemAttack. ϵ represents the perturbation amplitude.

E.2 Adversarial Defense Details

The following are the details of adversarial defense and audio transformations for Table 2 we use against semantic attacks:

Table 9: Hyperparameters of T-SemAttack.

Hyperparameter	Value
lr	$1e{-3}$
lr_min	$1e{-6}$
max_steps	500
ϵ	$8/255$

- **Resample.** Following standard adversarial defense practice [39], we randomly downsample the input waveform to one of 8000, 10000, 12000 Hz and then upsample it back to the original 24 kHz. This destroys high-frequency components where adversarial perturbations are often concentrated.
- **Quantize and Dequantize (Q-D).** The waveform is linearly mapped from the [-1, 1] range to a b-bit integer range and then mapped back to its original scale. This bit-depth reduction acts as a coarse projection that disrupts small adversarial perturbations in the waveform amplitude.
- **Mel-Inverse Reconstruction (Mel)**. The waveform is first transformed into a mel-spectrogram and then reconstructed back to the time domain using a Griffin–Lim–based inversion. This process is lossy and discards the original phase information, acting as a strong non-linear filter that removes perturbations not aligned with the mel-scale spectral structure. The configuration uses an FFT size of 2048, a hop length of 512, 128 mel bands, and a sampling rate of 24 kHz.
- **AudioPure** [31]. AudioPure is a defense method against adversarial audio attacks. It uses a diffusion model to purify perturbed audio by adding noise and then denoising it through the reverse diffusion process, reducing adversarial perturbations and making the audio closer to clean speech, thereby improving the robustness of speech recognition or audio classification systems under attack.
- **Change speed (Speed).** We apply time stretching with a randomly selected rate from the set 0.8, 0.9, 1.0, 1.1, 1.2. This temporally realigns acoustic events without altering pitch.
- **MP3 Compression (MP3).** We apply MP3 compression with bitrates uniformly sampled between 8 and 64 kbps while preserving the original waveform length. The compressed signal is either truncated or zero-padded as needed to maintain alignment.
- **Pitch Shift (PS).** We shift the pitch by a random number of semitones uniformly sampled from -4 to +4. This modifies the pitch contour while preserving the duration of the audio.
- **Band-Pass Filter (BPF).** We apply a band-pass filter with a center frequency uniformly sampled between 100 and 6000 Hz. This restricts the spectral content of the input, retaining only a specific frequency band.

The adversarial defense methods follow WaveGuard [39]. Audio transformation techniques are implemented using the audiomentations library

E.3 ROSETok and Training Details

Details of the ROSETok Architecture. The ROSETok framework comprises several core components: an encoder compresses the input audio waveform into a 1280-dimensional latent representation, which is then aligned with semantic embeddings from a Whisper Teacher model. A 9-codebook Residual Vector Quantizer subsequently discretizes this latent representation into compact codes, and a decoder reconstructs the waveform from these codes.

The DAC encoder and decoder are constructed from symmetrically arranged downsampling and upsampling modules. The encoder contains four EncoderBlocks that progressively downsample the signal with striding factors of 2, 4, 5, and 8, while expanding the channel dimension from 64 to 1024 and finally projecting to a 1280-dimensional latent space. Each EncoderBlock sequentially connects three residual units with dilation rates of 1, 3, and 9, followed by a Snake activation function. The decoder mirrors this architecture exactly, starting with 1536 channels and employing four DecoderBlocks with upsampling factors of 8, 5, 4, and 2 to restore the monaural audio waveform.

The quantizer adopts residual vector quantization with nine codebook layers, where each layer maintains 1024 entries of 8-dimensional vectors. The quantization dropout rate is set to zero to ensure training stability. Through successive residual refinement across codebook layers, a compact

8-dimensional discrete representation approximates the original 1280-dimensional continuous latent vector, achieving a compression ratio of approximately 17.8.

During the training stage, two additional components are introduced, namely the semantic encoder and the discriminator. During the training process, the semantic encoder employs a frozen Whisper large-v3 encoder, which first extracts Mel spectrogram features via short-time Fourier transform and then produces 1280-dimensional semantic embeddings through its Transformer backbone. This embedding dimensionality aligns precisely with the output dimension of the DAC encoder. Additionally, the discriminator is introduced to enhance the perceptual quality of the reconstructed audio by comparing it against the original audio, where adversarial loss and feature matching loss are computed to guide the training.

Details of Training Objectives. The details of each loss during training are as follows:

$\mathcal{L}_{\text{distill}}$ is the semantic distillation loss, which forces the audio encoder to learn semantically rich representations by maximizing the cosine similarity between the latent representations output by the encoder and the semantic embeddings extracted by the frozen Whisper teacher model [17]. The semantic distillation loss $\mathcal{L}_{\text{distill}}$ is computed as $\mathcal{L}_{\text{distill}} = 1 - \text{CosSim}(z, z_{\text{sem}})$, where z denotes the encoded representation and z_{sem} is the representation from the semantic encoder.

$\mathcal{L}_{\text{recon}}$ is the reconstruction loss, which ensures the fidelity of the codec process by minimizing the difference between the audio reconstructed by the decoder and the original audio.

$\mathcal{L}_{\text{cb}}$ is the codebook loss, which optimizes the codebook vectors of the residual vector quantizer to better fit the feature distribution output by the encoder.

$\mathcal{L}_{\text{commit}}$ is the commitment loss, which constrains the encoder output to stay close to the nearest codebook vector, preventing large fluctuations between different codebooks.

$\mathcal{L}_{\text{adv}}$ is the adversarial loss, which improves the perceptual quality of the generated audio by training a discriminator to distinguish between real and reconstructed audio.

$\mathcal{L}_{\text{fm}}$ is the feature matching loss, which stabilizes the adversarial training process and further enhances the quality of the reconstructed audio by minimizing the difference between features extracted from the original and reconstructed audio at intermediate layers of the discriminator.

We adopt $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{cb}}$, and $\mathcal{L}_{\text{commit}}$ from [40], and $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{fm}}$ from [21].

Details of Training Datasets. We train the ROSETok model on 96.7k hours of open-source English and Chinese data. All open-source datasets used in this work are listed in Table 10. LibriSpeech is used for ablation and the stage-two fine-tuning.

Table 10: Detailed information of training datasets.

Dataset	Subset	Duration (# h)	Task	Language	Usage
Librispeech	train & val	960	ASR	English	Ablation & ROSETok-AAF
Emilia	EN	46,800	ASR	English	ROSETok-Base
Emilia	ZH	49,900	ASR	Chinese	ROSETok-Base

Details of Hyperparameters. Table 11 summarizes the main hyperparameters used throughout ROSETok training. "lr" refers to the initial learning rate of the AdamW optimizer, controlling the step size of parameter updates during training. "lr_scheduler" refers to the learning rate scheduling strategy, where CosineLR is used to smoothly anneal the learning rate following a cosine curve. "batch_size" refers to the global number of audio samples per training step. "noise_level" refers to the noise amplitude during noisy training. "warmup_steps" refers to the initial 1000 steps during which the learning rate is linearly warmed up to stabilize early-stage training. "total_steps" refers to the total number of optimization steps $200{,}000$ for training the robust audio tokenizer.

Details of the XARES Benchmark. XARES [34] evaluates audio encoders through downstream probing tasks. We freeze the encoder, extract frame-level speech representations, and feed them into lightweight task-specific MLP heads for classification. Following the XARES protocol, raw task metrics are normalized to $[0, 1]$, where higher scores indicate stronger semantic capability. For readability, we report the normalized scores in percentage form.

Table 11: Training Hyperparameters of the ROSETok.

Hyperparameter	Value
ROSETok-Base	
optimizer	AdamW
lr	1.0×10^{-4}
lr_scheduler	CosineLR
min_lr	1.0×10^{-5}
batch_size	128
noise_level	$4/255 \sim 24/255$
warmup_steps	1000
total_steps	200,000
ROSETok-AAF	
optimizer	AdamW
lr	5.0×10^{-6}
lr_scheduler	CosineLR
min_lr	5.0×10^{-7}
batch_size	64
warmup_steps	100
total_steps	2,000

E.4 Baseline and Model Details

The baseline attack methods considered in our study are as follows: (1) **Gaussian Noise.** As the most fundamental comparison baseline, this method introduces perturbations by adding random Gaussian noise to the original audio signal. To ensure that the signal-to-noise ratio of the processed audio is comparable to that of our proposed method, facilitating a fair comparison, we set the noise standard deviation to 0.01. (2) **E2E-VGuard** [9] is an end-to-end adversarial defense scheme designed for speech models. In our experiments, we strictly follow the default parameter configurations from the original paper to ensure the reliability of reproduction results and the objectivity of the comparison. (3) **AudioShield** [14] attacks audio encoders by applying perturbations in the latent feature space, where a threshold parameter τ controls the trade-off between perturbation strength and perception. We select the τ = 0.1 to achieve a comparable perceptual results with T-SemAttack.

To comprehensively evaluate the effectiveness of our semantic attack, we test the adversarial audio against a diverse set of Automatic Speech Recognition (ASR) models spanning both open-source and closed-source systems. The open-source models we use are as follows:

(1) **HuBERT-Large** [16] is a self-supervised speech representation model fine-tuned on 960 hours of LibriSpeech data, known for its strong performance on English transcription tasks through masked prediction of hidden units. (2) **Whisper-Large** [17] is OpenAI's multilingual, multitask ASR model trained on 680k+ hours of weakly supervised audio data, offering robust performance across languages and acoustic conditions. HuBERT-Large and Whisper-Large are evaluated as ASR systems in Table 1, while their encoders are also used as surrogate branches during attack optimization. (3) **Qwen3-ASR-0.6B** [6] is Alibaba's lightweight ASR model built upon the Qwen3 foundation, providing efficient and accurate multilingual speech recognition with a compact 0.6B parameter footprint. (4) **VibeVoice-ASR** [24] is a recent ASR system designed for natural conversational speech, emphasizing robustness to spontaneous and expressive vocal styles. (5) **Fun-ASR-Nano** [26] is a compact variant of the FunASR family developed by Alibaba DAMO Academy, optimized for low-latency inference while maintaining competitive accuracy on Mandarin and multilingual benchmarks. (6) **FireRedASR** [25] is an attention-based encoder-decoder ASR model from Xiaohongshu, achieving SOTA results on Mandarin benchmarks with strong generalization to diverse acoustic scenarios.

For closed-source ASR systems, we evaluate against three widely-used commercial APIs: iFlytek, Doubao, and GPT-4o-mini-transcribe, which allows us to assess the transferability of our attack to black-box production systems. (1) **iFlytek** is a leading Chinese commercial ASR service renowned for its high accuracy on Mandarin and various dialects in industrial deployments. (2) **Doubao-ASR** is ByteDance's commercial speech recognition API, offering robust transcription capabilities

across multiple languages and noisy real-world conditions. (3) **GPT-4o-mini-transcribe** is OpenAI's lightweight transcription endpoint derived from the GPT-4o family, providing high-quality multilingual transcription through a unified multimodal architecture.

In addition to dedicated ASR systems, we further evaluate our attack against LALMs, which integrate audio understanding with large language model capabilities. (1) **Qwen2-Audio** [1] is Alibaba's instruction-tuned audio-language model that supports both voice chat and audio analysis, capable of processing diverse audio inputs including speech, sounds, and music. (2) **Qwen3-Omni** [2] is a flagship omni-modal model with a Mixture-of-Experts architecture (30B total / 3B active parameters), unifying text, image, audio, and video understanding with strong instruction-following capabilities. (3) **Audio-Flamingo-3** [5] is NVIDIA's advanced audio-language model that excels at understanding complex audio content, including speech, music, and ambient sounds, with strong reasoning and dialogue capabilities. The closed-source category includes **Gemini-3-Flash-Preview**, Google's latest multimodal model from the Gemini family, offering fast inference and strong cross-modal understanding across audio, vision, and text.

To ensure these models perform pure transcription rather than higher-level reasoning or summarization, we adopt a unified system prompt of "*Transcribe the audio exactly as recorded. Output the transcription only. Do not add explanations, summaries, or comments.*" This setup enables a fair comparison between LALMs and conventional ASR models under the same transcription objective.

To evaluate the effectiveness of our noise-robust tokenizer, we compare it against a diverse set of representative baseline tokenizers. (1) **SNAC** [41] is a multi-scale neural audio codec that employs hierarchical RVQ at different temporal resolutions to efficiently compress audio signals. (2) **XY-Tokenizer** [42] is a speech tokenizer designed to jointly model semantic and acoustic information, enabling unified representation for speech understanding and generation. (3) **XCodec-2** [21] integrates semantic features from self-supervised models with acoustic reconstruction objectives, producing discrete tokens that balance linguistic content and audio fidelity. (4) **DAC** [40] is a high-fidelity general-purpose neural audio codec that leverages improved RVQ techniques and adversarial training to achieve strong reconstruction quality across diverse audio domains. (5) **MOSS-Audio-Tokenizer** [43] is a neural audio codec that compresses raw audio into discrete tokens, enabling efficient audio generation and understanding within the MOSS multimodal framework. (6) **StableToken** [7] is a tokenizer that focuses on producing stable and consistent discrete representations across varying input conditions, aiming to mitigate token instability in downstream tasks. Together, these baselines cover a broad spectrum of design philosophies, ranging from purely acoustic codecs to semantically-enriched and stability-oriented tokenizers, providing a comprehensive baseline system for assessing noise robustness.

E.5 Compute Reporting and Efficiency Details

Details of Attacks' Efficiency Evaluation. In Table 1, the RTF is computed as the ratio of the per-sample perturbation generation time to the total duration of the audio in Eq. (6).

$$\mathrm{RTF}_{\mathrm{Attacks}} = \frac{\sum_{i=1}^{N} T_{\mathrm{attack}}^{(i)}}{\sum_{i=1}^{N} T_{\mathrm{audio}}^{(i)}}, \tag{6}$$

where $T_{\mathrm{attack}}^{(i)}$ is the time taken by each attack method to process the i-th audio sample, $T_{\mathrm{audio}}^{(i)}$ is the duration of the i-th audio sample, and N is the total number of test samples.

AudioShield constitutes a special case because it pre-trains a universal adversarial perturbation (UAP) on a training set, then directly applies it to all test inputs without per-audio optimization. Since the training stage is essential for obtaining a usable perturbation, we amortize its cost across test samples to ensure a fair comparison in Eq. (7).

$$\mathrm{RTF}_{\mathrm{AudioShield}} = \frac{T_{\mathrm{train}} + \sum_{i=1}^{N} T_{\mathrm{apply}}^{(i)}}{\sum_{i=1}^{N} T_{\mathrm{audio}}^{(i)}}, \tag{7}$$

where T_{train} is the total perturbation training time, $T_{\mathrm{apply}}^{(i)}$ is the time to apply the pre-trained perturbation to the i-th audio sample. This amortizes the training overhead across test samples, enabling a fair comparison with other baselines under a unified framework. When N is sufficiently large, the amortized T_{train} term becomes negligible, and the RTF of AudioShield gradually approaches its

theoretical lower bound. The RTF measurements of attacks are conducted on a device equipped with an NVIDIA 4090 GPU and Intel(R) Xeon(R) Platinum 8488C CPU, 512GB RAM.

Details of Tokenizers' Efficiency Evaluation. In Table 4, the Real-Time Factor (RTF) is computed as the ratio of the total audio reconstruction time to the total duration of the audio in Eq. (8).

$$\text{RTF}_{\text{Tokenizers}} = \frac{\sum_{i=1}^{N} T_{\text{reconstruct}}^{(i)}}{\sum_{i=1}^{N} T_{\text{audio}}^{(i)}}, \tag{8}$$

The reconstruction time $T_{\text{reconstruct}}^{(i)}$ measures only the tokenizer's core reconstruction process, namely waveform encoding and decoding. It excludes the time required to load audio files from disk or save reconstructed audio back to disk. This ensures that the reported RTF reflects the robust tokenizer's intrinsic computational efficiency, rather than I/O overhead unrelated to model performance. The RTF measurements of tokenizers are conducted on a device equipped with an NVIDIA 4090 GPU, Intel(R) Xeon(R) Platinum 8488C CPU, and 512 GB RAM.

Compute Reporting. We conduct our experiments using GPU-equipped computing resources. Specifically, we run T-SemAttack and the evaluation on a single NVIDIA RTX 4090 GPU. For training ROSETok, we use 8 NVIDIA H100 GPUs for approximately 20 hours.

F Visualization and Analyses of the T-SemAttack

F.1 Perturbation Amplification

To analyze the mechanism of T-SemAttack in the speech tokenizer, we compute the representation differences between clean audio x and adversarial audio x_{adv} at different internal stages of the tokenizer. We use the S3 Tokenizer as an example, and its audio processing pipeline is: Waveform $\rightarrow$ Mel-spectrogram $\rightarrow$ Encoder $\rightarrow$ VQ Tokens.

For continuous representation stages, we use relative L_2 distortion, referred to as Relative Distortion (RD), to measure the perturbation magnitude:

$$\text{RD}_s = \frac{\|z_s(x_{\text{adv}}) - z_s(x)\|_2}{\|z_s(x)\|_2},$$

where $z_s(\cdot) \in \mathbb{R}^{T_s \times d_s}$ denotes the representation at stage s, T_s is the sequence length of the representation at that stage, and d_s is the feature dimension. Note that T_s is stage-dependent and may differ from the waveform length T. We also use cosine distance (CD) to measure changes in the representation direction:

$$\text{CD}_s = 1 - \frac{1}{T_s} \sum_{t=1}^{T_s} \cos\left(z_{s,t}(x), z_{s,t}(x_{\text{adv}})\right).$$

The amplification factor at each stage is defined as the ratio between the average perturbation at that stage and the average perturbation at the waveform stage:

$$\text{Amp}_{\text{RD}}(s) = \frac{\frac{1}{N}\sum_{i=1}^{N} \text{RD}_s^{(i)}}{\frac{1}{N}\sum_{i=1}^{N} \text{RD}_{\text{wav}}^{(i)}}.$$

For the final discrete VQ tokens, we compute the token error rate (TER):

$$\text{TER} = 1 - \frac{\sum_{n=1}^{N} \mathbf{1}[c_n(x) = c_n(x_{\text{adv}})]}{N},$$

where $c_n(\cdot)$ denotes the n-th discrete token and N denotes the number of tokens.

Table 12 reports the perturbation amplification results on the S3 Tokenizer. The average RD at the waveform stage is only 0.154, but it rapidly increases to 0.771 at the Mel-spectrogram stage, corresponding to a $5.01\times$ amplification. After the encoder, the average RD further increases to 1.520, which is $9.87\times$ larger than that at the waveform stage. Meanwhile, the average CD of the encoder output reaches 1.248, indicating a severe directional shift in the high-level semantic representations of the adversarial samples compared with the clean samples.

Table 12: Perturbation amplification results in the S3 Tokenizer.

Stage	RD Mean	RD Std	RD Median	CD Mean	CD Median	Amp(RD)
Waveform	0.154	0.023	0.153	0.322	0.326	$1.00\times$
Mel Spectrogram	0.771	0.070	0.772	0.357	0.347	$5.01\times$
Encoder Output	1.520	0.055	1.530	1.248	1.250	$9.87\times$

Finally, at the VQ token level, T-SemAttack achieves an average TER of 99.7%. This shows that perturbation drift in the continuous representation space is ultimately quantized into an almost entirely different discrete token sequence. Therefore, T-SemAttack does not merely introduce local waveform noise. Instead, it creates a progressive amplification chain inside the S3 Tokenizer, from the acoustic frontend to the semantic encoder, ultimately causing tokenizer-level semantic collapse.

F.2 LALM Attention

To further analyze the impact of T-SemAttack on the cross-modal modeling process of LALMs, we take Qwen2-Audio as an example and conduct a layer-wise analysis of Transformer attention under clean and adversarial audio inputs. The analysis consists of a single-sample attention-map visualization, which intuitively shows changes in the attention distribution before and after the attack.

Single-Sample Attention Visualization. To illustrate attention disruption, we select one clean-adversarial pair and visualize the attention maps of several representative layers. For each layer, we average over all attention heads and then plot the resulting two-dimensional attention matrix. In the figure, the x-axis denotes key positions, and the y-axis denotes query positions. Since the LLM component of Qwen2-Audio uses causal attention, the future-token region in the upper-right corner is masked, appearing as a black triangular area.

Figure 8 presents a comparison between clean and adversarial attention maps. Brighter colors correspond to larger attention weights from the query token to the key token. To display both strong and weak attention regions, we use log-scale normalization in the visualization.

To better observe the attention changes caused by the attack, we further plot the differential attention map as $\Delta A_l = A_l^{\mathrm{adv}} - A_l^{\mathrm{clean}}$. Figure 9 shows the differential maps across different layers. Red indicates increased attention under the adversarial condition, while blue indicates decreased attention.

The differential maps show that attention changes are more pronounced in the middle and upper layers, and are concentrated near the audio boundaries, the target-text start position, and several key positions strongly relied on by the model.

Therefore, the single-sample visualization provide complementary visualizations. Figure 8 and Figure 9 illustrate how this change appears in specific samples, layers, and token regions. These observations suggest that once the semantic representations of the speech tokenizer are disrupted by our semantic attack, the cross-modal alignment process of the LALM is also affected, making it difficult for the model to recover the correct semantics from the audio context and ultimately leading to understanding degradation.

F.3 Transcription Examples.

We provide some attack examples in Table 13.

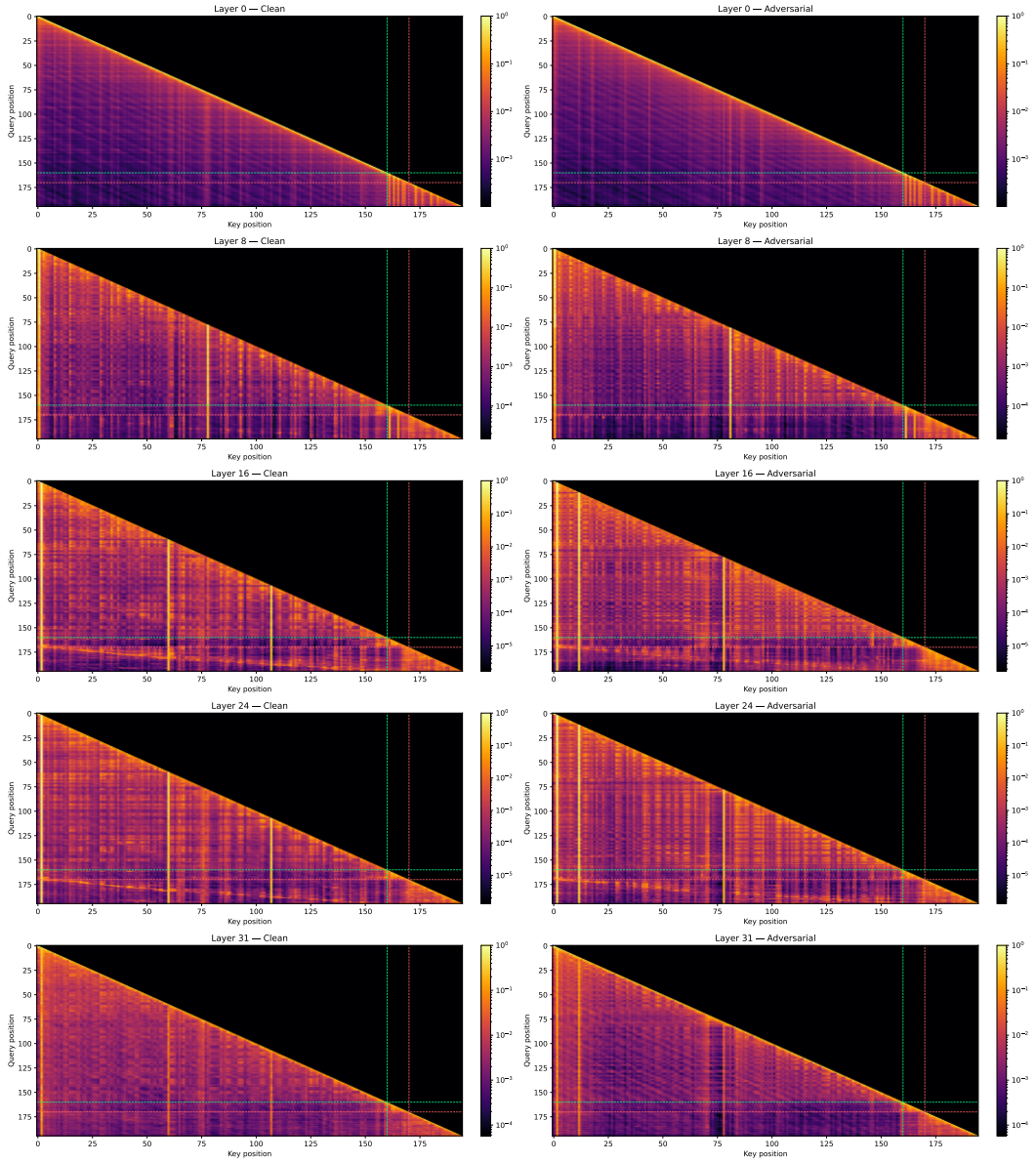


Figure 8: Single-sample attention maps of Qwen2-Audio under clean and adversarial inputs. Each row corresponds to one Transformer layer, while the left and right columns show clean and adversarial maps, respectively. Green dashed lines mark the audio-token boundaries, and red dashed lines mark the beginning of the target-text region.

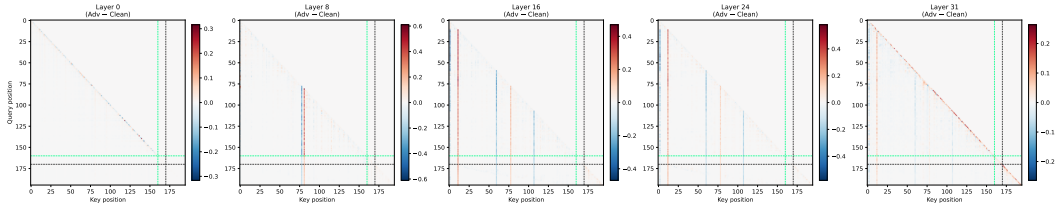


Figure 9: Single-sample attention difference maps between adversarial and clean inputs. The visualized value is $\Delta A_l = A_l^{\mathrm{adv}} - A_l^{\mathrm{clean}}$. Red regions indicate increased attention under adversarial input, while blue regions indicate decreased attention.

Table 13: Transcriptions on different ASR models

Method	WER (%)	Transcriptions
Whisper Large		
Original	–	roaming endlessly around the park she wants to go home
Clean	0.000	roaming endlessly around the park she wants to go home
Gaussian	10.000	blooming endlessly around the park she wants to go home
E2E-VGuard	10.000	roaming endlessly around the park he wants to go home
AudioShield	90.000	roman did not listen to the wrong that hour she wanted to go home
T-SemAttack	250.000	and olympics and are on the path to a free world home have fun watching people ? have fun watching people ? have fun watching people ?
VibeVoice-ASR		
Original	0.000	i am never more aware of a room is acoustics than when i am trying to enjoy a snack i have no intention of sharing
Gaussian	12.000	i am never more aware of a room is acoustics than when i am trying to listen back i have no intention of sharing
E2E-VGuard	64.000	i am never going to really believe the postscript and i am calling you just now i have no intention to
AudioShield	28.000	i am never more aware of a woman is acoustics than when i am trying to enjoy a snack i have not looked into the tree already
T-SemAttack	116.000	i never know where to leave i do not think it is going to be fun it is going to be good i hope you do not think it is too late
Qwen3-ASR		
Original	0.000	i am never more aware of a room is acoustics than when i am trying to enjoy a snack i have no intention of sharing
clean	4.000	i am never more aware of a room is coziest than when i am trying to enjoy a snack i have no intention of sharing
Gaussian	4.000	i am never more aware of a room is coasticks than when i am trying to enjoy a snack i have no intention of sharing
E2E-VGuard	44.000	i am never more aware of a wind is approach than when one gives me a snack i have made them so care
AudioShield	28.000	i am never more aware of romanticostics than when trying to eat a snack i have no intention of sharing
T-SemAttack	88.000	i remember we are going to coast beach in lantau the green map i have used on the computer

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are supported by the proposed T-SemAttack, the analysis of tokenizer-level semantic fragility, and the ROSETok defense. The scope is reflected in our evaluations on multiple datasets, 9 ASR systems, 4 LALMs, adaptive defenses, and robust tokenizer reconstruction and downstream-tasks experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Appendix A, including perceptual imperceptibility versus normal usability and offline efficiency.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not present formal theoretical results, theorems, or proofs. The paper focuses on a semantic-based attack, mechanistic analysis, and a robust tokenizer design.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the attack objective, optimization process, surrogate encoders, psychoacoustic constraint, datasets, evaluation metrics, target systems, and tokenizer training details in Sections 3, 6, and Appendix E. These details include hyperparameters, optimization iterations, training data, and compute settings. Moreover, our utilized datasets are open-sourced.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments are conducted on publicly available benchmark datasets and accessible open-source or commercial systems. The code link is available at `https://github.com/wxzyd123/T-SemAttack`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 specifies the datasets, sampled subsets, ASR/LALM systems, metrics, and ROSETok training hyperparameters. Appendix E further provides implementation details such as training objectives, dataset duration, optimizer settings, RTF computation, and hardware.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: For the subjective human listening study, we report MOS scores with 95% confidence intervals to reduce the influence of subjective bias. For attack and reconstruction evaluations, we report averaged values over the selected evaluation sets.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For attack evaluation, we conducted experiments on an NVIDIA 4090 GPU. For ROSETok training and evaluation, we disclose the compute resources on 8 NVIDIA H100 GPUs and RTF measurements on an NVIDIA 4090 GPU with an Intel Xeon CPU and 512GB RAM.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We considered the NeurIPS Code of Ethics, especially because T-SemAttack is a dual-use technique. Appendix B discusses misuse risks, third-party system usage, human-subject protections, and responsible release considerations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix B discusses both positive impacts, such as exposing tokenizer vulnerabilities and motivating robust speech tokenizers, and negative impacts, such as misuse for degrading ASR, evading moderation, or interfering with voice assistants and accessibility systems.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The T-SemAttack research implementation is publicly available. The repository documents intended use and misuse risks, and Appendix B discusses the dual-use implications.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the datasets, models, baselines, and third-party systems used in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new research artifacts, including T-SemAttack implementation resources and ROSETok-related code or model components. These assets are documented with usage instructions, intended-use statements, limitations, and responsible-use notes in our paper and GitHub link.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix D describes the subjective listening study, including recruitment through Credamo, participant requirements, sample construction, MOS scoring dimensions, $1.47 per participant as the compensation, the full text of instructions given to participants and screenshots are shown in Fig. 7.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The subjective survey was approved by the Human Ethics Research Committee at the first author's institution. We obtained informed consent, recruited only participants over 18 years old, anonymized participant identities, and used responses only for academic research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.