JOINT ATTENTION-SEMANTIC DISRUPTION ATTACK ON NON-AUTOREGRESSIVE TEXT-TO-SPEECH MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in neural speech synthesis have shifted toward non-autoregressive models, but the adversarial robustness of such continuous-trajectory and non-token-based schemes remains under-explored. In this paper, we reveal that the mapping between the continuous acoustic prompt and the generation trajectory is highly susceptible to subtle cross-modal perturbations. We exploit this vulnerability and propose a novel black-box adversarial attack that simultaneously distorts acoustic embeddings, dissolves internal cross-modal alignment, and corrupts semantic representations. Extensive experiments show that the proposed attack successfully sabotages five state-of-the-art text-to-speech models with a small ℓ_∞ perturbation budget $\epsilon = 0.01$ (SNR$\approx$ 19dB, audio waveform scaled to $[-1, 1]$) under eight representative adversarial defense measures: total variation regularization, Kullback-Leibler regularization, 100% adversarial training, low-pass filtering, μ-law companding, randomized smoothing, the Griffin-Lim algorithm, and score-based speech enhancement. Our method can be applied to low-energy digital jamming, availability attack, and evaluation of generative audio robustness.

1 INTRODUCTION

The text-to-speech (TTS) technology has achieved significant development due to the rapid breakthroughs in neural audio synthesis. State-of-the-art (SOTA) systems ranging from flow-matching models to scalable zero-shot frameworks can synthesize natural and high-fidelity human speech from only seconds of reference audio. This capability has inspired widespread commercial deployment in virtual assistants (Le et al., 2023; Chen et al., 2025a), dynamic content creation (Tan, 2023; Ju et al., 2024), and real-time translation (Communication et al., 2023; Jia et al., 2022). However, the quick expansion of these TTS models has gone far beyond our understanding of their structural vulnerabilities, which yields an urgent need to rigorously evaluate their security boundaries against unauthorized exploitation.

Deep neural networks (DNNs) are shown to be vulnerable to imperceptible and worst-case input perturbations (Goodfellow et al., 2015; Carlini & Wagner, 2017; Madry et al., 2018). In the context of audio, extensive research focuses on automatic speech recognition (ASR) and automatic speaker verification (ASV) pipelines (Carlini & Wagner, 2018; Lea et al., 2019; Wu et al., 2021; Wang et al., 2024). Specific audio defense mechanisms include randomized smoothing (Cohen et al., 2019), low-pass filtering (Sun et al., 2020), μ-law companding (Hsu & Liao, 2012), the Griffin-Lim algorithm (Nenov et al., 2024), and score-based speech enhancement (Richter et al., 2023). Other adversarial defense measures are Total Variation (TV) regularization (Rudin et al., 1992; Lai & Wang, 2024; Wang et al., 2025a; Lai et al., 2026a;b), Kullback-Leibler (KL) regularization (Goldblum et al., 2020; Zhang et al., 2025), and adversarial training (AT, Xie et al. 2019; Shafahi et al. 2019). However, as we will show later, these defense mechanisms still fail under specially-developed adversarial attacks that aim at the inherent vulnerabilities of TTS models. Therefore, the security problem of generative audio systems remains under-explored.

In this work, we identify and exploit several **common vulnerabilities** of non-autoregressive (NAR) TTS models: flow-matching (Lipman et al., 2023), the zero-shot setting, sensitive cross-attention

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

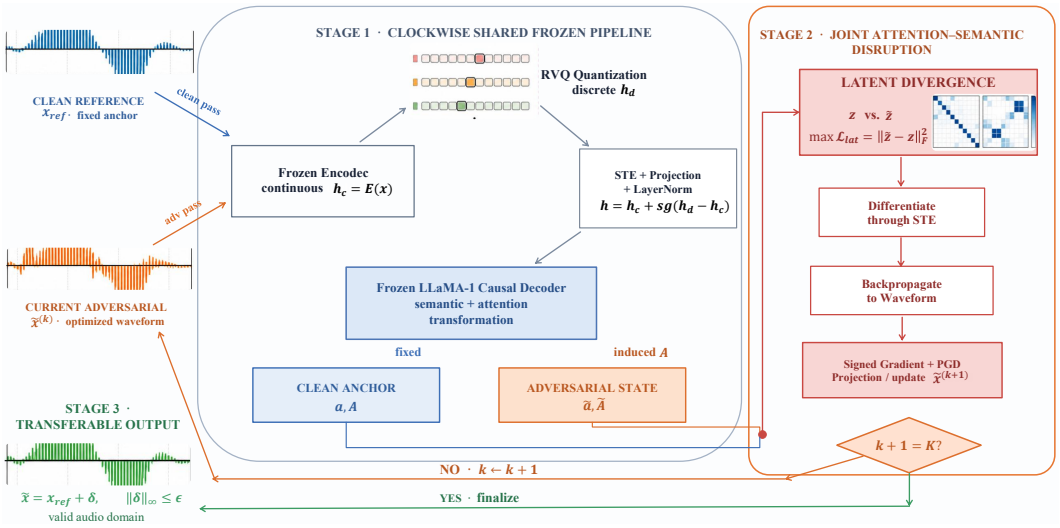


Figure 1: Attack scheme of the proposed JASDA. It iteratively maximizes the latent divergence between the clean reference audio and the adversarial audio via a shared and frozen structure including `Encodec` and `Llama1`.

modules, and discrete token detonation. **1)** Flow matching relies on highly stable and straight-line ordinary differential equation (ODE) trajectories to transform noise into clear speech. An early-stage semantic or acoustic deviation injected by the adversarial sample may compound linearly across the integration steps and result in distorted or complete catastrophic noise synthesis. **2)** In the zero-shot setting, NAR TTS models rely heavily on one single brief reference audio prompt to extract speaker identity, channel characteristics, and acoustic environments via high-capacity speech encoders (Chen et al., 2025a; Wang et al., 2025b), which is vulnerable to latent feature hijacking. **3)** The sensitive cross-attention modules project the adversarial perturbations directly into the latent workspace, which may be faithfully replicated throughout the entire generated output. **4)** Since such TTS models sample tokens based on the joint probability distribution of past choices, a small early adversarial perturbation may trigger a runaway chain reaction that steers these models away from the correct context.

Based on the above vulnerabilities, we propose a Joint Attention-Semantic Disruption Attack (**JASDA**), which simultaneously achieves the following disruptive effects:

1. It forces the bottleneck layers of the target system to misinterpret the voice identity of a speaker by maximizing the Mean Squared Error (MSE) between the latent representations of the clean and the perturbed samples.

2. It isolates the gradient flow strictly to the adversarial path, which avoids the vanishing problem caused by the zero-derivatives of the quantization function.

3. It forces semantic collapse across different speech architectures by maximizing the divergence between the latent representations of the clean and perturbed reference audio.

JASDA requires only two common tools in audio processing: `Encodec` (Défossez et al., 2023) and `Llama1` (Touvron et al., 2023). Only a single application of JASDA to a clean reference waveform to produce a slightly perturbed waveform is enough for the attack (Figure 1). The attacker has no access to and only minimal knowledge on the target TTS model, thus mainstream black-box attack schemes like the square attack (Andriushchenko et al., 2020) are not applicable. Once the perturbed waveform is input into the target TTS model, the corresponding output waveform deteriorates sharply in the speaker similarity score (SIM) (Figure 2), which is a primary evaluating metric for speaker verification. JASDA successfully sabotages five SOTA NAR TTS models: F5-TTS (Chen et al., 2025b), E2-TTS (Eskimez et al., 2024), MaskGCT (Wang et al., 2025b), ZipVoice (Zhu et al., 2025), and LEMAS-TTS (Zhao et al., 2026), even under the eight representative defense methods mentioned above. It reveals that the underlying geometry of such speech generation models remains vulnerable to structural attacks.

JASDA can be applied to at least three practical scenarios:

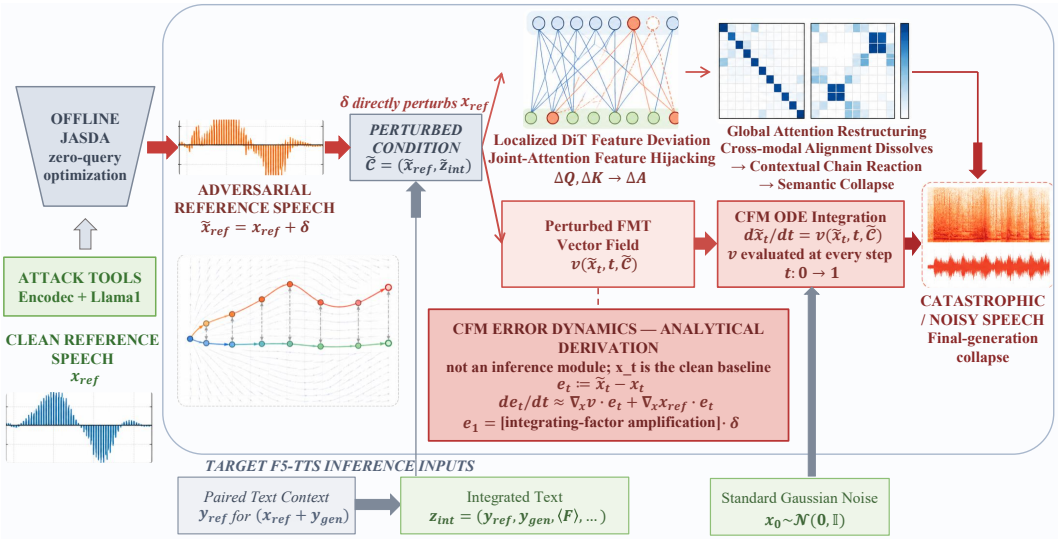


Figure 2: Attack chain of the proposed JASDA on F5-TTS. It perturbs clean reference speech to trigger joint-attention feature hijacking and vector field distortion, which amplifies errors through CFM ODE integration to cause semantic collapse and generate catastrophic noise.

1. It can act as an efficient digital jammer that protects voice privacy or restricts unauthorized speech synthesis pipelines by scrambling the generation process before it even starts.
2. It can render the target NAR TTS model unavailable by corrupting its output with an unnoticeable perturbed reference audio clip.
3. It can serve as a rigorous diagnostic tool to evaluate the structural security boundaries of new speech synthesis architectures and build more robust generative audio models.

2 Related Works

2.1 Non-autoregressive Text-to-speech Models

Modern NAR TTS models decouple or parallelize the generation process to further enable parallel generation. A typical approach is the flow-matching, which constructs a continuous vector field $v_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ to define an ODE that morphs a simple source distribution, such as the standard normal distribution $p(x) = \mathcal{N}(x|0, \mathbb{I})$, into the target acoustic trajectory x_t: $\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\theta(x_t, t, \mathcal{C})$, where $\mathcal{C}$ represents some given condition with respect to (w.r.t.) this inference. This model is optimized via a Conditional Flow Matching (CFM, Lipman et al. 2023) objective over the time horizon $t \sim \mathcal{U}(0, 1)$:

$$\begin{aligned}\mathcal{L}_{CFM}(\theta) &= \mathbb{E}_{t,q(x_1),p(x_0)} \| v_\theta(\psi_t(x_0), t) - \frac{\mathrm{d}}{\mathrm{d}t}\psi_t(x_0) \|_2^2 \\ &= \mathbb{E}_{t,q(x_1),p(x_0)} \| v_\theta((1-t)x_0 + tx_1, t) - (x_1 - x_0) \|_2^2, \end{aligned} \tag{1}$$

where θ contains the trainable parameters, and $\psi_t(x_0)$ is the deterministic mapping trajectory between noise and data. Since there is no prior knowledge, a conditional probability $p_t(x|x_1) = \mathcal{N}(x|\mu_t(x_1), \sigma_t(x_1)^2\mathbb{I})$ can be used in training to obtain identical gradients w.r.t. θ. Then the flow map $\psi_t(x) = \sigma_t(x_1)x + \mu_t(x_1)$ with initialization $\sigma_0(x_1) = 1$ and $\mu_0(x_1) = 0$. At the end of the flow, $\sigma_1(x_1) = 0$ and $\mu_1(x_1) = x_1$ are set to enforce convergence of the path to p_1. With the reparameterization technique (Kingma & Welling, 2014; Rezende et al., 2014), the left equality of (1) holds. Besides, using the optimal transport form (Lipman et al., 2023) $\psi_t(x) = (1-t)x + tx_1$ leads to the right equality of (1).

Different NAR TTS models have different architectures of $v_\theta(\cdot, t)$. We take E2-TTS and F5-TTS as examples, and let $x_1 \in \mathbb{R}^{F \times N}$ be an extracted mel spectrogram feature vector for the speech example x to be used in training, where F and N denote the mel dimension and the sequence length, respectively. Then the noisy speech $(1-t)x_0 + tx_1$, the masked speech $(1-\mathcal{M}) \odot x_1$, and the training target $\mathcal{M} \odot (x_1 - x_0)$ will be used in (1) later, where $\mathcal{M} \in \{0, 1\}^{F \times N}$ denotes a binary temporal mask. Besides, the transcript y corresponding to x is extended to

$\mathfrak{z} = (\underbrace{c_1, c_2, \cdots, c_M}_{y}, \underbrace{\langle\mathcal{F}\rangle, \cdots, \langle\mathcal{F}\rangle}_{(N-M)times})$, where c_i denotes the i-th character of y and $\langle\mathcal{F}\rangle$ denotes a filler token with the same length as one mel frame. The training process is to reconstruct $\mathcal{M} \odot x_1$ with $(1 - \mathcal{M}) \odot x_1$ and $\mathfrak{z}$. In the inference stage, the following form of v_t is used:

$$v_\theta(\psi_t(x_0), t, \mathcal{C}) = v_\theta((1-t)x_0 + tx_1, t, x_{ref}, \mathfrak{z}_{int}), \tag{2}$$

where x_{ref} and $\mathfrak{z}_{int} = (y_{ref}, y_{gen}, \langle\mathcal{F}\rangle, \cdots)$ denote the reference speech and the text integrated by the reference transcript y_{ref} and the prompt y_{gen}, respectively. Therefore, x_{ref} and $\mathfrak{z}_{int}$ serve as the inference condition $\mathcal{C}$. We will show later that only a **small adversarial perturbation** to x_{ref} can **significantly degrade** the performance of target NAR TTS models.

2.2 Adversarial Attack

In the perspective of this paper, an adversarial attack aims to sabotage target NAR TTS models by corrupting a clean reference speech $x := x_{ref}$ with an imperceptible perturbation δ. The resulting adversarial example $\tilde{x} = x + \delta$ is tailored to induce verification failures (we denote the **perturbed version** of a variable by adding a **tilde** over it in the rest of this paper). Formally, the attacker optimizes a task-specific loss function $\mathcal{L}$ to maximize the prediction error:

$$\max_{\|\delta\|_\mathfrak{p} \leqslant \epsilon} \mathcal{L}(x + \delta), \tag{3}$$

where the perturbation magnitude is bounded by a predefined *budget* $\epsilon > 0$ under an $\ell_\mathfrak{p}$-norm constraint. In a white-box circumstance, the attacker possesses full knowledge of the target architecture and its loss landscape, thus $\mathcal{L}$ can be optimized directly. Representative algorithms in this category include the Fast Gradient Sign Method (FGSM, Goodfellow et al. 2015), Projected Gradient Descent (PGD, Madry et al. 2018; Croce & Hein 2020a), C&W (Carlini & Wagner, 2017), Auto PGD (APGD), and AutoAttack (Croce & Hein, 2020b). Among them, PGD serves as a representative iterative approach that updates the perturbation via:

$$\tilde{x}^{(k+1)} = \mathrm{proj}_{\mathcal{B}_\mathfrak{p}[x;\epsilon]}(\tilde{x}^{(k)} + \eta \cdot \mathrm{sign}(\nabla_x \mathcal{L}(\tilde{x}^{(k)}))), \tag{4}$$

where $\eta > 0$ denotes the step size, and $\mathrm{proj}_{\mathcal{B}_\mathfrak{p}[x;\epsilon]}$ represents the projection onto the closed $\ell_\mathfrak{p}$-ball centered at x with radius ϵ.

On the other hand, conventional black-box techniques typically rely on querying the model to obtain outputs $y = f(x)$ (Papernot et al., 2017). A standard approach involves using historical input-output pairs $\{(x, y)\}$ to train a surrogate network $\hat{f}$, which is then used to generate adversarial examples. **In contrast, our threat model operates under a more restricted circumstance: it requires zero interaction with the target model and executes in a single-shot fashion, where query-based or surrogate-based black-box attacks are inapplicable.**

3 Methodology

In this section, we present how to construct an adversarial reference speech $\tilde{x}_{ref}$ via JASDA, and demonstrate the corresponding attack chain on target NAR TTS models. JASDA requires only two common tools for necessary audio processing: `Encodec` and `Llama1`. The former is a high-fidelity neural network based audio compressor that converts continuous raw audio waves into tokens, while the latter is a self-supervised learning framework for speech transcriptions and representations. By exploiting both tools, JASDA can simultaneously manipulate speech tokens and disrupt the phonetic understanding of target models (see Figure 1).

3.1 Latent Representation and Attention Extraction

At the beginning of the attack, JASDA initializes the adversarial speech with the clean reference speech $\tilde{x}^{(0)} := x_{ref}$, and subsequently optimizes it over sufficient iterations. At the k-th attack iteration, the first operation is to encode the current adversarial speech $\tilde{x}^{(k)}$. Specifically, suppose $\tilde{x}^{(k)} \in \mathbb{R}^T$ with both batch size and channel size being 1, where T represents the raw temporal dimension. Using the Encodec $\mathcal{E}$ with the default pretrained parameters θ_{enc} on $\tilde{x}^{(k)}$ yields a continuous latent representation:

$$\tilde{h}_c^{(k)} = \mathcal{E}(\tilde{x}^{(k)}; \theta_{enc}) \in \mathbb{R}^{D \times T'}, \tag{5}$$

where D is the latent dimensionality, and $T' = \lceil \frac{T}{S} \rceil$ is the downsampled temporal length determined by the total stride S of the convolutional layers. The next step is to quantize $\tilde{h}_c^{(k)}$ into a discrete form

$\tilde{h}_d^{(k)}$ with bandwidth $\mathcal{B}$, N_q discrete codebooks, $\mathscr{K}$ vectors for each codebook, and dimensionality= D for each vector. The relationships between these terms are:

$$\mathfrak{f}_{fr} = \frac{\mathfrak{f}_s}{S}, \quad N_q = \frac{1000 \cdot \mathcal{B}}{\mathfrak{f}_{fr} \cdot \log_2(\mathscr{K})}, \tag{6}$$

where $\mathfrak{f}_s$ and $\mathfrak{f}_{fr}$ denote the sampling and the frame frequencies, respectively. For a common setting, $\mathfrak{f}_s = 24000$ $(24kHz)$, $S = 320$, $\mathcal{B} = 6$, and $\mathscr{K} = 1024$. Then calculations from (6) lead to $\mathfrak{f}_{fr} = 75Hz$ and $N_q = 8$. Let $e_{j,\mathfrak{k}} \in \mathbb{R}^D$ be the $\mathfrak{k}$-th vector of the j-th codebook, the residual vector quantization (RVQ, Défossez et al. 2023) yields:

$$\begin{aligned} &R_1(t) = \tilde{h}_c^{(k)}(t); \quad \mathfrak{y}_j(t) := \underset{\mathfrak{k}\in\{1,\cdots,\mathscr{K}\}}{\arg\min} \|R_j(t) - e_{j,\mathfrak{k}}\|_2^2, \\ &R_{j+1}(t) = R_j(t) - e_{j,\mathfrak{y}_j(t)}, j = 1, \cdots, N_q; \quad \tilde{h}_d^{(k)}(t) = \sum_{j=1}^{N_q} e_{j,\mathfrak{y}_j(t)}, t = 1, \cdots, T', \end{aligned} \tag{7}$$

where $\mathfrak{y}_j(t)$ is the discrete token index selected via a nearest-neighbor search in the current codebook space with ℓ_2 norm.

Since adversarial perturbations are typically minor, their gradient information might be eliminated during the backpropagation due to the zero-derivatives of the quantization function. Motivated by the Straight-Through Estimator (STE, Bengio et al. 2013), we develop the following backpropagation step to avoid the above vanishing problem:

$$\tilde{h}^{(k)} = \tilde{h}_c^{(k)} + \mathrm{sg}(\tilde{h}_d^{(k)} - \tilde{h}_c^{(k)}), \tag{8}$$

where $\mathrm{sg}(\cdot)$ is the *stop gradient* operator that *detaches* the model parameter update from the argument. It creates a bifurcated relationship during the execution flow:

$$\text{Forward Pass:} \quad \tilde{h}^{(k)} = \tilde{h}_d^{(k)}; \quad \text{Backpropagation:} \quad \nabla_x \mathcal{L} = \nabla_x h \cdot \nabla_h \mathcal{L}, \nabla_h \mathcal{L} = \nabla_{h_c} \mathcal{L}. \tag{9}$$

Note that **previous works** (Bengio et al., 2013; Défossez et al., 2023) exploit STE mainly to **train model parameters** like θ_{enc}, which involves $\nabla_\theta \mathcal{L}$; but **JASDA** uses STE to **construct an adversarial example**, which involves $\nabla_x \mathcal{L}$.

The next step is to feed the audio features $\tilde{h}^{(k)}$ into `Llama1`, in order to interpret what the audio means. Specifically, the input for `Llama1` is initialized via an affine transform followed by a $(0, 1)$-layer normalization $\mathcal{T}_{LN}$:

$$\tilde{z}_{(0)}^{(k)} := \mathcal{T}_{LN}((\tilde{h}^{(k)})^\top W_{\text{proj}} + b_{\text{proj}}) \in \mathbb{R}^{T' \times D_{LLM}}, \quad W_{\text{proj}} \in \mathbb{R}^{D \times D_{LLM}}, b_{\text{proj}} \in \mathbb{R}^{D_{LLM}}, \tag{10}$$

where W_{proj} and b_{proj} are the weight and the bias matrices for the affine transform. (10) maps the speech representation $\tilde{h}^{(k)}$ from the acoustic space into the linguistic space of `Llama1`. Then $\tilde{z}_{(0)}$ passes through L consecutive Transformer layers of `Llama1`. For each layer $l \in [1, L]$, the Multi-Head Self-Attention (MHSA) calculates the Query, Key, and Value matrices per head $m \in \{1, \cdots, M\}$:

$$Q_{(l),m}^{(k)} = \mathcal{T}_{RN}(\tilde{z}_{(l-1)}^{(k)}) W_{(l),m}^Q, K_{(l),m}^{(k)} = \mathcal{T}_{RN}(\tilde{z}_{(l-1)}^{(k)}) W_{(l),m}^K, V_{(l),m}^{(k)} = \mathcal{T}_{RN}(\tilde{z}_{(l-1)}^{(k)}) W_{(l),m}^V, \tag{11}$$

where $\mathcal{T}_{RN}$ denotes the Root mean square Normalization (RN, Zhang & Sennrich 2019). To further exploit the relative position dependency in self-attention, the rotary positional embeddings (RoPE, Su et al. 2024) are added to each layer:

$$\tilde{Q}_{(l),m}^{(k)} = Q_{(l),m}^{(k)} \mathcal{R}_{\theta_{rot},m}, \tilde{K}_{(l),m}^{(k)} = K_{(l),m}^{(k)} \mathcal{R}_{\theta_{rot},m}, \tag{12}$$

where $\mathcal{R}_{\theta_{rot},m}$ denotes the rotary matrix for the m-th head with parameters θ_{rot} pre-defined as in (Su et al., 2024). Then the attention score matrix $\tilde{A}_{(l),m}^{(k)} \in \mathbb{R}^{T' \times T'}$ can be calculated by

$$\tilde{A}_{(l),m}^{(k)} = \mathrm{softmax}\left(\frac{\tilde{Q}_{(l),m}^{(k)} (\tilde{K}_{(l),m}^{(k)})^\top}{\sqrt{\mathfrak{d}_m}} + \mathcal{M}_{cau} \right), \quad [\mathcal{M}_{cau}]_{ij} := \begin{cases} 0 & \text{if } j \leqslant i \\ -\infty & \text{if } j > i \end{cases}, \tag{13}$$

where $\mathfrak{d}_m$ denotes the dimensionality of the m-th head, and $\mathcal{M}_{cau}$ denotes the causal attention mask matrix that imposes temporal causality on the attention score. Next, $\tilde{A}^{(k)}_{(l),m}$ routes information from the latent representations to construct the next state layer:

$$\tilde{z}^{(k)}_{(l-1)} = \tilde{z}^{(k)}_{(l-1)} + (\|_{m=1}^{M}(\tilde{A}^{(k)}_{(l),m}\mathcal{T}_{RN}(\tilde{z}^{(k)}_{(l-1)})W^{V}_{(l),m}))W^{O}_{(l)},$$
$$\tilde{z}^{(k)}_{(l)} = \tilde{z}^{(k)}_{(l-1)} + \mathcal{T}_{FFN}(\mathcal{T}_{RN}(\tilde{z}^{(k)}_{(l-1)})), \tag{14}$$

where $W^{O}_{(l)}$ and $\mathcal{T}_{FFN}$ denote the attention Output matrix and the Feed-Forward Network (FFN) with the SwiGLU activation (Shazeer, 2020) for `Llama1`, respectively. $\|_{m=1}^{M}$ concatenates the dominating matrices along the attention head dimension.

3.2 Joint Attention-Semantic Disruption

After L layers, the last hidden state of the adversarial sample is produced $\tilde{z}^{(k)}_{(L)} \in \mathbb{R}^{T' \times D_{LLM}}$, directly denoted by $\tilde{z}^{(k)}$ for simplicity. On the other hand, its clean counterpart also has the last latent state z_{ref}. Then JASDA aims to cast a joint attention-semantic disruption by maximizing the squared Frobenius distance between $\tilde{z}^{(k)}$ and z_{ref}:

$$\mathcal{L}^{(k)}_{lat}(\tilde{x}^{(k)}) := \|\tilde{z}^{(k)} - z_{ref}\|^2_F := \|z(\tilde{x}^{(k)}) - z(x_{ref})\|^2_F, \tag{15}$$

where $z(\cdot)$ is written as the mapping from the speech domain to the latent domain for subsequent notational convenience. According to the deductions from (11) to (14), we can find that the latent representations $\{\tilde{z}^{(k)}_{(l)}\}$ is coupled with the attentions $\{\tilde{A}^{(k)}_{(l),m}\}$. Hence maximizing $\mathcal{L}^{(k)}_{lat}$ jointly disrupts both $\{\tilde{z}^{(k)}_{(l)}\}$ and $\{\tilde{A}^{(k)}_{(l),m}\}$, which would be made clear later.

Next, we calculate the partial gradient $\partial_x \mathcal{L}^{(k)}_{lat}(\tilde{x}^{(k)})$, in order to find an adversarial update direction. According to the chain rule,

$$\mathcal{J}_l := [\partial_{z_{(l-1)}} z_{(l)}]_{\text{dir}} + \sum_{m=1}^{M} \partial_{A^{(k)}_{(l),m}} z_{(l)} \cdot \partial_{z_{(l-1)}} A^{(k)}_{(l),m}, \partial_{h_c} h = \underbrace{\partial_{h_c} h_c}_{\mathbb{I}} + \underbrace{\partial_{h_c} \text{sg}(h^{(k)}_d - h^{(k)}_c)}_{0}, \tag{16}$$

$$\partial_{h_c} z_{(0)} = \partial_h z_{(0)} \cdot \partial_{h_c} h = \partial_h z_{(0)}, \partial_x \mathcal{L}^{(k)}_{lat}|_{x=\tilde{x}^{(k)}} = \partial_{z_{(L)}} \mathcal{L}^{(k)}_{lat} \cdot \left(\prod_{l=L}^{1} \mathcal{J}_l\right) \cdot \partial_h z_{(0)} \cdot \partial_x h_c|_{x=\tilde{x}^{(k)}}, \tag{17}$$

where $[\cdot]_{\text{dir}}$ denotes the direct gradient flowing through the literal latent representations with the attention maps frozen. Then the adversarial speech at the $(k+1)$-th attack iteration can be updated by one-step input optimization (Erhan et al., 2009; Goodfellow et al., 2015; Madry et al., 2018; Gurumurthy et al., 2021; Yu & Gao, 2024):

$$\tilde{x}^{(k+1)} = \text{proj}_{\mathcal{B}_\infty[x_{ref};\epsilon] \cap \mathcal{X}}(\tilde{x}^{(k)} + \eta \cdot \text{sign}(\partial_x \mathcal{L}^{(k)}_{lat}(\tilde{x}^{(k)}))), \tag{18}$$

where the operator $\text{proj}_{\mathcal{B}_\infty[x_{ref};\epsilon] \cap \mathcal{X}}$ projects the output speech onto the intersection of the speech domain $\mathcal{X}$ and the closed ball centered at the clean reference speech x_{ref} within an ℓ_∞ distance of ϵ. In practice, a speech is usually normalized into $\mathcal{X} = [-1, 1]^{\mathfrak{d}_x}$ for the Encodec, where $\mathfrak{d}_x$ denotes the dimensionality of x_{ref}. After $\mathcal{K}$ attack iterations, the final perturbed speech $\tilde{x} := \tilde{x}^{(K)}$ is constructed.

JASDA significantly differs from existing mainstream attacks in the following aspects. **1)** Unlike query-dependent black-box methods like the Square Attack (Andriushchenko et al., 2020), JASDA operates in a zero-query and single-shot manner without requiring target model interaction. It targets universal acoustic-semantic invariants via independent foundation tools, which enables adversarial transferability across unknown NAR TTS models. **2)** While white-box attacks like PGD and C&W optimize perturbations against specific task boundaries, JASDA bypasses quantization constraints via an STE to maximize representation divergence directly in the latent self-attention space. **3)** Instead of manipulating text strings like standard prompt evasion, JASDA injects an imperceptible perturbation into the raw reference audio prompt. Since zero-shot NAR models rely heavily on this reference clip to establish the generation context, poisoning it triggers a runaway chain reaction that corrupts the entire output speech.

3.3 ATTACK CHAIN

We take F5-TTS as an example to illustrate the attack chain of JASDA (see Figure 2), which is a SOTA NAR TTS model built upon a Flow Matching Transformer (FMT) architecture. It involves several vulnerable modules that could be targeted: CFM, joint-attention features, and contextual chain reaction.

Linear error amplification in CFM. CFM solves ODEs to map standard Gaussian noise into mel-spectrogram trajectories, which expects a stable straight-line vector field. The adversarial perturbation δ of JASDA disrupts the initial conditions of ODEs, and this early-stage acoustic deviation compounds linearly across successive integration steps, which enforces the ODE solver to drift off-course and synthesize catastrophic noise. Specifically, denote $e_t := \tilde{x}_t - x_t$ as the trajectory error vector at time step t. Applying a first-order multivariate Taylor expansion to the perturbed vector field $v(\tilde{x}_t, t, \tilde{\mathcal{C}})$ around the clean state $(x_t, \mathcal{C})$ with $\mathcal{C} = (x_{ref}, \mathfrak{z}_{int})$ defined in (2) yields:

$$v(\tilde{x}_t, t, \tilde{\mathcal{C}}) \approx v(x_t, t, \mathcal{C}) + \nabla_x v(x_t, t, \mathcal{C}) e_t + \nabla_{x_{ref}} v(x_t, t, \mathcal{C}) \delta, \tag{19}$$

$$\frac{\mathrm{d}e_t}{\mathrm{d}t} = v(\tilde{x}_t, t, \tilde{\mathcal{C}}) - v(x_t, t, \mathcal{C}) \approx \nabla_x v(x_t, t, \mathcal{C}) e_t + \nabla_{x_{ref}} v(x_t, t, \mathcal{C}) \delta, \tag{20}$$

where $\nabla_x v$ and $\nabla_{x_{ref}} v$ denote derivatives w.r.t. the internal evolving state x and the external conditioning input x_{ref}, respectively. Note that x_t changes continuously at every integration step, while x_{ref} remains constant over the generation time $t \in [0, 1]$. $\nabla_x v_t$ determines how the system behaves internally once the noise is inside, while $\nabla_{x_{ref}} v_t$ determines how much external noise enters the generation pipeline. By introducing a state-transition matrix $\Phi(\tau, s)$, the analytical solution for the accumulated error e_t at the final synthesis boundary $t = 1$ can be calculated as:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\Phi(\tau, s) = \nabla_x v(x_\tau, \tau, \mathcal{C})\, \Phi(\tau, s), \quad \forall \tau \in [0, 1], \quad \text{with initial condition } \Phi(s, s) = I,$$

$$e_1 = \underbrace{\Phi(1, 0) e(0)}_{=0} + \underline{\left(\int_0^1 \Phi(1, s) \nabla_{x_{ref}} v(x_s, s, \mathcal{C})\, \mathrm{d}s \right) \cdot \delta}, \tag{21}$$

which indicates that the adversarial perturbation δ can be **linearly amplified** with the underlined factor. Because $\Phi(\tau, s)$ generally belongs to the exponential function class, even a localized and brief period of **positive divergence** during the trajectory will cause the error propagator to **scale exponentially**.

Joint-attention feature hijacking. The joint multi-modal attention mechanisms in F5-TTS propagate continuous acoustic perturbations across the shared sequence length, and cause the final generation to collapse. Specifically, instead of passing through a discrete bottleneck or an isolated audio encoder, the adversarial perturbation δ directly affects the continuous reference speech input x_{ref}: $\tilde{\mathcal{C}} = (\tilde{x}_{ref}, \tilde{\mathfrak{z}}_{int})$. This triggers a localized feature deviation in the multi-modal Diffusion Transformer (DiT) blocks via the first-order Taylor expansion (19). Then the perturbed conditioning input goes through a **similar forward pass** as (10)$\sim$(14) (but with a **different attention mechanism and model structure**: `Llama1` vs. DiT), which dissolves the internal cross-modal alignment with the text embedding as shown in (15).

Contextual chain reaction. Because joint text-audio tokens are decoded contextually based on the global prompt environment, the localized adversarial perturbation of JASDA can trigger a semantic collapse. Once F5-TTS misinterprets the primary phonetic dependencies, the perturbation would steer the neural synthesizer away from the correct context and invalidate F5-TTS.

Theorem 1. *According to* (13)*, the attention differential can be calculated as follows:*

$$\mathrm{d}A_{(l),m} := \frac{1}{\sqrt{\mathfrak{d}_m}} A_{(l),m} \odot [\underbrace{(\mathrm{d}Q_{(l),m} K_{(l),m}^\top + Q_{(l),m}\, \mathrm{d}K_{(l),m}^\top)}_{\textit{Core Driven Shift}}$$

$$- \mathrm{rowsums}\Big(A_{(l),m} \odot \underbrace{(\mathrm{d}Q_{(l),m} K_{(l),m}^\top + Q_{(l),m}\, \mathrm{d}K_{(l),m}^\top)}_{\textit{Core Driven Shift}} \Big) \mathbb{1}^\top], \tag{22}$$

where $\odot$ *and* rowsums *denote the Hadamard product and the rowsum operator, respectively.* $\mathbb{1} \in \mathbb{R}^{T' \times 1}$ *is the matrix with all elements being* 1.

The proof is given in Appendix A.1. Let $\Delta A_{(l),m} := \tilde{A}_{(l),m} - A_{(l),m}$, $\Delta Q_{(l),m} := \tilde{Q}_{(l),m} - Q_{(l),m}$, and $\Delta K_{(l),m} := \tilde{K}_{(l),m} - K_{(l),m}$, then $\Delta A_{(l),m}$ can be approximated by $\Delta Q_{(l),m}$ and $\Delta K_{(l),m}$ via (22). It indicates how an adversarial perturbation propagates through the non-linear Softmax layer of F5-TTS, and how localized changes in latent features translate into a global restructuring of the attention weights. In particular, the **core driven shift** in (22) represents the **first-order change** in the unnormalized and pre-activation attention scores, which is split into two independent pathways where the **displacement** of either query or key is **amplified** by the **current vectors** of each other. This term dictates how it inadvertently projects its attention onto other tokens (e.g., text tokens) under the JASDA attack. Attack chains on other target NAR TTS models are provided in Appendix A.2.

4 Experimental Results

The experimental setup strictly follows the standard benchmark in the literature of TTS. Since the proposed JASDA attack mainly targets the similarity between the generated and the reference audios, the SIM metric (see Appendix A.4.3) is used as the main evaluating metric. We report SIM scores for the clean cases, the attacked cases, and the attacked plus defended cases for all the five target models and eight defense measures. These target models and defense measures are detailed in Section 3.3, Appendices A.2 and A.3. Three mainstream English speech datasets are used: LibriSpeech-PC (Meister et al., 2023), Emilia (He et al., 2024), and Common Voice (Ardila et al., 2020). Detailed experimental settings are presented in Appendix A.4. We also provide **additional experiments in Appendix A.5, including results on two datasets, ablation analysis, and hyperparameter sensitivity analysis**.

4.1 Main Results on LibriSpeech-PC

Table 1 presents the SIM scores of five SOTA NAR TTS models under the JASDA attack on the LibriSpeech-PC dataset across eight representative defense strategies. Overall, JASDA shows significant adversarial transferability and effectiveness in a black-box setting, which inflicts severe performance deterioration across all target architectures.

Vulnerability of continuous trajectory models. Flow-matching and continuous-trajectory architectures (F5-TTS and E2-TTS) suffer the most significant degradation. Without defense, JASDA reduces the SIM of F5-TTS from a clean 0.7736 to 0.2647 ($\Delta\text{SIM} = -0.5089$, a 65.78% RSD, see Eq. 82 and 83) and E2-TTS from 0.7814 to 0.3114 ($\Delta\text{SIM} = -0.4700$, a 60.15% RSD). This vulnerability aligns with our theoretical analysis: because continuous flow-matching relies on straight-line ODE integration trajectories, early-stage perturbations injected by JASDA into the reference audio compound linearly over time steps, which steer the generated acoustic trajectory off-course into catastrophic output noise.

Impact on masked and hybrid architectures. ZipVoice and LEMAS-TTS also experience substantial quality loss under undefended attack, with RSDs of 40.13% and 46.05%, respectively. MaskGCT shows relatively higher resilience under the baseline attack (23.27% RSD, falling from 0.6546 to 0.5023), owing to its discrete token masking mechanism. However, when combined with the SSE defense, MaskGCT suffers severe performance collapse with its SIM plunging to 0.3149 (51.89% RSD).

Ineffectiveness of existing defenses. Standard adversarial defenses fail to restore speaker similarity against JASDA:

- *Adversarial training and regularization (AT, KL, TV):* 100% AT and KL regularization provide little defense for flow-matching models (e.g., F5-TTS retains a 66.33% RSD under AT) and even exacerbate degradation in LEMAS-TTS (SIM dropping further to 0.3353, a 56.74% RSD).
- *Signal-level and time-domain defenses (μ-law, RS, LPF, GL):* Non-linear companding (μ-law) and RS leave the attack impact nearly unchanged (SIM changes < 0.01). LPF provides slight reduction, recovering F5-TTS SIM to 0.3889 and E2-TTS to 0.4537, but RSDs remain high (49.73% and 41.94%).
- *Generative purification (SSE):* While SSE moderately improves SIM for LEMAS-TTS (0.5008) and ZipVoice (0.4929), it remains far below clean baselines and unexpectedly triggers catastrophic disruption on MaskGCT (a 51.89% RSD).

Table 1: SIM scores for five target models under the JASDA attack and eight defense measures on the LibriSpeech-PC dataset. Each defense cell reports SIM, paired ΔSIM, and RSD.

Defense	ZipVoice	LEMAS-TTS	MaskGCT	F5-TTS	E2-TTS
Clean	SIM **0.7461**	SIM **0.7750**	SIM **0.6546**	SIM **0.7736**	SIM **0.7814**
No Defense	SIM **0.4467** $\Delta\text{SIM}_{\text{atk}}$ -0.2994 RSD 40.13%	SIM **0.4181** $\Delta\text{SIM}_{\text{atk}}$ -0.3569 RSD 46.05%	SIM **0.5023** $\Delta\text{SIM}_{\text{atk}}$ -0.1523 RSD 23.27%	SIM **0.2647** $\Delta\text{SIM}_{\text{atk}}$ -0.5089 RSD 65.78%	SIM **0.3114** $\Delta\text{SIM}_{\text{atk}}$ -0.4700 RSD 60.15%
AT	SIM **0.4453** $\Delta\text{SIM}_{\text{def}}$ -0.0014 RSD 40.32%	SIM **0.3353** $\Delta\text{SIM}_{\text{def}}$ -0.0828 RSD 56.74%	SIM **0.5190** $\Delta\text{SIM}_{\text{def}}$ +0.0167 RSD 20.71%	SIM **0.2605** $\Delta\text{SIM}_{\text{def}}$ -0.0042 RSD 66.33%	SIM **0.2825** $\Delta\text{SIM}_{\text{def}}$ -0.0289 RSD 63.85%
KL	SIM **0.4156** $\Delta\text{SIM}_{\text{def}}$ -0.0311 RSD 44.30%	SIM **0.3549** $\Delta\text{SIM}_{\text{def}}$ -0.0632 RSD 54.21%	SIM **0.5147** $\Delta\text{SIM}_{\text{def}}$ +0.0124 RSD 21.37%	SIM **0.2472** $\Delta\text{SIM}_{\text{def}}$ -0.0175 RSD 68.05%	SIM **0.3559** $\Delta\text{SIM}_{\text{def}}$ +0.0445 RSD 54.45%
TV	SIM **0.4694** $\Delta\text{SIM}_{\text{def}}$ +0.0227 RSD 37.09%	SIM **0.4621** $\Delta\text{SIM}_{\text{def}}$ +0.0440 RSD 40.37%	SIM **0.5157** $\Delta\text{SIM}_{\text{def}}$ +0.0134 RSD 21.22%	SIM **0.3352** $\Delta\text{SIM}_{\text{def}}$ +0.0705 RSD 56.67%	SIM **0.3408** $\Delta\text{SIM}_{\text{def}}$ +0.0294 RSD 56.39%
GL	SIM **0.4784** $\Delta\text{SIM}_{\text{def}}$ +0.0317 RSD 35.88%	SIM **0.4545** $\Delta\text{SIM}_{\text{def}}$ +0.0364 RSD 41.35%	SIM **0.5040** $\Delta\text{SIM}_{\text{def}}$ +0.0017 RSD 23.01%	SIM **0.3387** $\Delta\text{SIM}_{\text{def}}$ +0.0740 RSD 56.22%	SIM **0.3581** $\Delta\text{SIM}_{\text{def}}$ +0.0467 RSD 54.17%
μ-law	SIM **0.4479** $\Delta\text{SIM}_{\text{def}}$ +0.0012 RSD 39.97%	SIM **0.4183** $\Delta\text{SIM}_{\text{def}}$ +0.0002 RSD 46.03%	SIM **0.4943** $\Delta\text{SIM}_{\text{def}}$ -0.0080 RSD 24.49%	SIM **0.2642** $\Delta\text{SIM}_{\text{def}}$ -0.0005 RSD 65.85%	SIM **0.3109** $\Delta\text{SIM}_{\text{def}}$ -0.0005 RSD 60.21%
LPF	SIM **0.4953** $\Delta\text{SIM}_{\text{def}}$ +0.0486 RSD 33.61%	SIM **0.4790** $\Delta\text{SIM}_{\text{def}}$ +0.0609 RSD 38.19%	SIM **0.5405** $\Delta\text{SIM}_{\text{def}}$ +0.0382 RSD 17.43%	SIM **0.3889** $\Delta\text{SIM}_{\text{def}}$ +0.1242 RSD 49.73%	SIM **0.4537** $\Delta\text{SIM}_{\text{def}}$ +0.1423 RSD 41.94%
RS	SIM **0.4460** $\Delta\text{SIM}_{\text{def}}$ -0.0007 RSD 40.22%	SIM **0.4114** $\Delta\text{SIM}_{\text{def}}$ -0.0067 RSD 46.92%	SIM **0.5091** $\Delta\text{SIM}_{\text{def}}$ +0.0068 RSD 22.23%	SIM **0.2618** $\Delta\text{SIM}_{\text{def}}$ -0.0029 RSD 66.16%	SIM **0.3064** $\Delta\text{SIM}_{\text{def}}$ -0.0050 RSD 60.79%
SSE	SIM **0.4929** $\Delta\text{SIM}_{\text{def}}$ +0.0462 RSD 33.94%	SIM **0.5008** $\Delta\text{SIM}_{\text{def}}$ +0.0827 RSD 35.38%	SIM **0.3149** $\Delta\text{SIM}_{\text{def}}$ -0.1874 RSD 51.89%	SIM **0.3099** $\Delta\text{SIM}_{\text{def}}$ +0.0452 RSD 59.94%	SIM **0.3517** $\Delta\text{SIM}_{\text{def}}$ +0.0403 RSD 54.99%

Main results on the Emilia and Common Voice datasets are provided in Appendices A.5.1 and A.5.2, respectively. All these results confirm that by simultaneously targeting latent representations and cross-modal attention maps via shared foundation encoders, JASDA creates widely effective and structurally entrenched adversarial perturbations that bypass both spatial regularization and signal purification.

5 Conclusion

In this paper, we develop the Joint Attention-semantic Disruption Attack (JASDA), a novel black-box adversarial attack that reveals deep structural vulnerabilities in non-autoregressive (NAR) text-to-speech (TTS) models. Operating under a single-shot and zero-query threat model, JASDA exploits foundation tools like `Encodec` and `Llama1` to disrupt continuous acoustic reference prompts without requiring direct access to target model architectures, parameters, or output queries. By employing a Straight-Through Estimator (STE) to overcome quantization gradient vanishing, JASDA simultaneously scrambles latent speaker embeddings, dissolves cross-modal attention alignment, and triggers semantic collapse.

Through comprehensive experiments, we show that a minor ℓ_∞ perturbation budget ($\epsilon = 0.01$, SNR $\approx$ 19dB) significantly corrupts the speaker similarity of five state-of-the-art NAR TTS systems. Moreover, eight representative defense strategies fail to restore generation quality or defend against the cross-modal disruption of JASDA. Our findings highlight the urgent need to strengthen the underlying security boundaries of continuous-trajectory and prompt-dependent generative audio architectures. In addition to serving as a diagnostic evaluation tool, JASDA can be applied to voice privacy protection, low-energy digital jamming, and model availability assessment. Future work includes designing secure cross-modal attention mechanisms and resilient ODE trajectory integration solvers capable of resisting latent structural perturbations.

REFERENCES

eSpeak NG (version 1.52), 2024. URL https://github.com/espeak-ng/espeak-ng.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pp. 484–501, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58591-4.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Stephen Butterworth et al. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Nicholas Carlini and David Wagner. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text . In *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7, Los Alamitos, CA, USA, May 2018. IEEE Computer Society.

Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1):89–97, 2004.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025a.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, Vienna, Austria, July 2025b. Association for Computational Linguistics.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 09–15 Jun 2019.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. Seamlessm4t: Massively multilingual & multimodal machine translation, 2023.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2196–2205. PMLR, 13–18 Jul 2020a.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 2206–2216, 2020b.

Nauman Dawalatabad, Mirco Ravanelli, François Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na. Ecapa-tdnn embeddings for speaker diarization. In *Proceedings of the Interspeech 2021*, pp. 3560–3564, 2021.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 682–689, 2024.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3996–4003, Apr. 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 369–376, New York, NY, USA, 2006. Association for Computing Machinery.

Swaminathan Gurumurthy, Shaojie Bai, Zachary Manchester, and J. Zico Kolter. Joint inference and input optimization in equilibrium networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16818–16832. Curran Associates, Inc., 2021.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890, 2024.

Chau-Yun Hsu and Hsuan-Chun Liao. Papr reduction using the combination of precoding and mu-law companding techniques for ofdm systems. In *2012 IEEE 11th International Conference on Signal Processing*, volume 1, pp. 1–4, 2012.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10120–10134. PMLR, 17–23 Jul 2022.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22605–22623. PMLR, 21–27 Jul 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Zhao-Rong Lai and Weiwen Wang. Invariant risk minimization is a total variation model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 25913–25935. PMLR, 21–27 Jul 2024.

Zhao-Rong Lai, Xiwen Yuan, Ziliang Chen, Liangda Fang, and Yongsen Zheng. A unified total variation framework for membrane potential perturbation dynamic. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Zhao-Rong Lai, Xiwen Yuan, and Jian Weng. Adversarial attack and defense for denoising diffusion sampling. In *The 43rd International Conference on Machine Learning*, 2026b.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 14005–14034. Curran Associates, Inc., 2023.

Schönherr Lea, Kohls Katharina, Zeiler Steffen, Holz Thorsten, and Kolossa Dorothea. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *Proceedings 2019 Network and Distributed System Security Symposium*, 2019.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.

Rossen Nenov, Dang-Khoa Nguyen, Peter Balazs, and Radu Ioan Boţ. Accelerated griffin-lim algorithm: A fast and provably converging numerical method for phase retrieval. *IEEE Transactions on Signal Processing*, 72:190–202, 2024.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pp. 506–519, New York, NY, USA, 2017. Association for Computing Machinery.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Ha Nguyen, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaëlle Laperrière, Mickael Rouvier, Renato De Mori, and Yannick Estève. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25 (333), 2024.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Noam Shazeer. Glu variants improve transformer, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312.

Ke Sun, Chen Chen, and Xinyu Zhang. "alexa, stop spying on me!": speech privacy protection against voice assistants. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, SenSys '20, pp. 298–311, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375900.

Xu Tan. *Neural text-to-speech synthesis*. Springer, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Kun Wang, Xiangyu Xu, Li Lu, Zhongjie Ba, Feng Lin, and Kui Ren. FraudWhistler: A resilient, robust and plug-and-play adversarial example detection method for speaker recognition. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 7303–7320, Philadelphia, PA, August 2024. USENIX Association. ISBN 978-1-939133-44-1.

Yuanchao Wang, Zhao-Rong Lai, and Tianqi Zhong. Out-of-distribution generalization for total variation based invariant risk minimization. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, June 2023.

Haibin Wu, Xu Li, Andy T. Liu, Zhiyong Wu, Helen Meng, and Hung-yi Lee. Adversarial defense for automatic speaker verification by cascaded self-supervised learning models. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6718–6722, 2021.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. Zipformer: A faster and better encoder for automatic speech recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

Hongzhan Yu and Sicun Gao. Activation-descent regularization for input optimization of relu networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57441–57458. PMLR, 2024.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Xu Zhang, Kaidi Xu, Ziqing Hu, and Ren Wang. Optimizing robustness and accuracy in mixture of experts: A dual-model approach. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76436–76450. PMLR, 13–19 Jul 2025.

Zhiyuan Zhao, Lijian Lin, Ye Zhu, Kai Xie, Yunfei Liu, and Yu Li. Lemas: Large a 150k-hour large-scale extensible multilingual audio suite with generative speech models, 2026.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025.

A Appendix

A.1 Proof of Theorem 1

Proof. The key step of the proof is to derive the first-order total differential of a row-wise Softmax matrix. For clarity, we omit the layer and head indices $(l), m$, and let the pre-activation matrix be $B := \frac{QK^\top}{\sqrt{\mathfrak{d}_m}} + \mathcal{M}_{cau}$. Then the attention matrix is defined element-wise as:

$$A_{ij} = \frac{\exp(B_{ij})}{\sum_{k=1}^{T} \exp(B_{ik})}. \tag{23}$$

Taking the total differential $\mathrm{d}A_{ij}$ with respect to the matrix elements $\mathrm{d}B$ yields:

$$\mathrm{d}A_{ij} = \frac{\exp(B_{ij})\,\mathrm{d}B_{ij}\,(\sum_k \exp(B_{ik})) - \exp(B_{ij})\,(\sum_k \exp(B_{ik})\,\mathrm{d}B_{ik})}{(\sum_k \exp(B_{ik}))^2}. \tag{24}$$

Factoring out terms to transform this in terms of the original probabilities A_{ij} and A_{ik} yields:

$$\begin{aligned}
\mathrm{d}A_{ij} =& \left(\frac{\exp(B_{ij})}{\sum_k \exp(B_{ik})}\right) \mathrm{d}B_{ij} - \left(\frac{\exp(B_{ij})}{\sum_k \exp(B_{ik})}\right) \sum_{k=1}^{T} \left(\frac{\exp(B_{ik})}{\sum_r \exp(B_{ir})}\right) \mathrm{d}B_{ik} \\
=& A_{ij}\,\mathrm{d}B_{ij} - A_{ij} \sum_{k=1}^{T} A_{ik}\,\mathrm{d}B_{ik} \\
=& A_{ij} \left(\mathrm{d}B_{ij} - \sum_{k=1}^{T} A_{ik}\,\mathrm{d}B_{ik}\right).
\end{aligned} \tag{25}$$

Now we translate this element-wise notation into a global matrix form with the Hadamard (element-wise) product $\odot$. Let $\mathbb{1} \in \mathbb{R}^{T\times 1}$ be a column vector of ones. The summation term $\sum_k A_{ik}\,\mathrm{d}B_{ik}$ represents the row-wise sum of the matrix $(A \odot \mathrm{d}B)$. We can express the row-sum broadcasting operation across all columns as $\mathrm{rowsums}(A \odot \mathrm{d}B)\mathbb{1}^\top$. Therefore,

$$\mathrm{d}A = A \odot \left(\mathrm{d}B - \mathrm{rowsums}(A \odot \mathrm{d}B)\mathbb{1}^\top\right). \tag{26}$$

Next, we expand the differential of the scaling operator $\mathrm{d}B$:

$$\mathrm{d}B = \mathrm{d}\left(\frac{QK^\top}{\sqrt{\mathfrak{d}_m}} + \mathcal{M}_{cau}\right) = \frac{1}{\sqrt{\mathfrak{d}_m}}\left(\mathrm{d}QK^\top + Q\,\mathrm{d}K^\top\right). \tag{27}$$

Substituting (27) back into (26) gives (22). □

A.2 Attack Chains on Other Target Non-autoregressive Text-to-speech Models

A.2.1 Attack Chain on LEMAS-TTS

LEMAS-TTS extends the training scheme of F5-TTS by exploiting a unified phonetic input space (eSpeak-NG, esp 2024), auxiliary supervision objectives (Connectionist Temporal Classification, CTC, Graves et al. 2006), and fine-grained prosody modeling (Seamless Expressive architecture, Communication et al. 2023). Hence LEMAS-TTS not only shares similar vulnerabilities to those of F5-TTS in Section 3.3, but also introduces some new vulnerabilities due to the above mechanisms.

Phonetic-level bypass space. Using eSpeak-NG cannot defend against JASDA, because JASDA attacks the reference speech instead of the prompt. Let $\mathscr{T}$ be the discrete space of natural language text tokens for eSpeak-NG, then the adversarial sample shares the same text token as that of the clean sample:

$$\tau_{adv} = \tau_{cln}, \quad \tau_{adv}, \tau_{cln} \in \mathscr{T}. \tag{28}$$

A text-based safety filter is defined as a scoring classifier $\mathscr{S}_{\mathscr{T}} : \mathscr{T} \to [0, 1]$ with a safety threshold $\mathfrak{t}$. A text string τ is rejected if $\mathscr{S}_{\mathscr{T}}(\tau) > \mathfrak{t}$. Since $\tau_{adv} = \tau_{cln}$, we have $\mathscr{S}_{\mathscr{T}}(\tau_{adv}) = \mathscr{S}_{\mathscr{T}}(\tau_{cln})$. Hence the adversarial sample has the same rejecting status as that of the clean sample.

CTC alignment sabotage and path collapse. The text-audio synchronization relies on a frame-level latent matrix $\tilde{z} = z + \delta$ generated by the DiT, which is mapped to phoneme posteriors via a linear projection layer W_{CTC}. Denote $\mathcal{V}' := \mathcal{V} \cup \{\mathfrak{e}\}$ as the extended phoneme vocabulary with the blank token $\mathfrak{e}$. The frame-level emission probability for any token $\mathfrak{k} \in \mathcal{V}'$ at frame t in the clean case is:

$$p(\mathfrak{k} \mid z(t)) = \frac{\exp(W_{\text{CTC},\mathfrak{k}}^\top z(t))}{\sum_{j\in\mathcal{V}'} \exp(W_{\text{CTC},j}^\top z(t))}. \tag{29}$$

The perturbation δ may skew this distribution at some tokens in some frame set Γ:

$$\begin{aligned} p(\mathfrak{k}' \mid \tilde{z}(t)) =& \frac{\exp(W_{\text{CTC},\mathfrak{k}'}^\top \tilde{z}(t))}{\sum_{j\in\mathcal{V}'} \exp(W_{\text{CTC},j}^\top \tilde{z}(t))} \neq \frac{\exp(W_{\text{CTC},\mathfrak{k}'}^\top z(t))}{\sum_{j\in\mathcal{V}'} \exp(W_{\text{CTC},j}^\top z(t))}, \\ & \exists\, \mathfrak{k}' \in \mathcal{V}', \quad \forall\, t \in \Gamma \subseteq [1, T]. \end{aligned} \tag{30}$$

Latent space explosion via Gradient Reversal Layer (GRL) collapse. The structural GRL creates an unconstrained conditioning space $\mathscr{C}$. We can map out how δ breaks the speaker identity boundary by evaluating the Singular Value Decomposition (SVD) of the Cumulative Condition Sensitivity Matrix $\mathscr{M}$:

$$\mathscr{M} = \int_0^1 \Phi(1, s) J_{\mathcal{C}}(s)\, \mathrm{d}s, \tag{31}$$

where $\Phi(1, s)$ denotes the state-transition matrix that describes the fundamental solution of the homogeneous system $\frac{d\Phi}{dt} = J_z(t)\Phi$:

$$\frac{d}{dt}\Phi(t, s) = J_z(t)\Phi(t, s), \quad \Phi(s, s) = I. \tag{32}$$

$J_{\mathcal{C}}$ and J_z denote the Jacobians of the LEMAS vector field v_{LEM} w.r.t. the conditioning $\mathcal{C}$ and the latent representation z, respectively. Using the variation of parameters, the error at the final execution frame $t = 1$ can be explicitly solved as:

$$e(1) = \Phi(1, 0)e(0) + \int_0^1 \Phi(1, s) J_{\mathcal{C}}(s)\delta\, \mathrm{d}s = \mathscr{M}\delta, \tag{33}$$

since $e(0) = 0$. If the vector field v_{LEM} is sensitive to the conditioning $\mathcal{C}$ (i.e., a large $\|J_{\mathcal{C}}\|_2$), the final error will amplify the perturbation δ with a large factor $\mathscr{M}$.

A.2.2 Attack Chain on E2-TTS

Compared with F5-TTS, E2-TTS has no text encoder or refinement block like ConvNeXt V2 (Woo et al., 2023). The character sequence padded with filler token and the acoustic mel-spectrogram features are entangled into a single joint matrix sequence. Without loss of generality and omitting indices $(l), m$ for simplicity,

$$\tilde{z} := [\tilde{x}_{ref} \| \tilde{\mathfrak{z}}_{int}], \tilde{Q} = \tilde{z}W^Q, \tilde{K} = \tilde{z}W^K, \tilde{B} = \frac{\tilde{Q}\tilde{K}^\top}{\sqrt{\mathfrak{d}_m}}, \tilde{A}_{ij} = \frac{\exp(\tilde{B}_{ij})}{\sum_{k=1}^T \exp(\tilde{B}_{ik})}. \tag{34}$$

Following similar deductions to those in Appendix A.1, we have:

$$\mathrm{d}B = \frac{1}{\sqrt{\mathfrak{d}_m}} \left(\mathrm{d}Q\tilde{K}^\top + \tilde{Q}\, \mathrm{d}K^\top \right), \tag{35}$$

$$\begin{aligned} \mathrm{d}A =& \tilde{A} \odot \left(\mathrm{d}B - \mathrm{rowsums}(\tilde{A} \odot \mathrm{d}B)\mathbb{1}^\top \right) \\ =& \frac{1}{\sqrt{\mathfrak{d}_m}} \tilde{A} \odot \left(\left(\mathrm{d}Q\tilde{K}^\top + \tilde{Q}\, \mathrm{d}K^\top \right) - \mathrm{rowsums}(\tilde{A} \odot \left(\mathrm{d}Q\tilde{K}^\top + \tilde{Q}\, \mathrm{d}K^\top \right))\mathbb{1}^\top \right). \end{aligned} \tag{36}$$

This is the same as the attention divergence result in Theorem 1. Besides, JASDA can also cause linear error amplification and hijack joint-attention features in E2-TTS, which are similar to those in Section 3.3.

A.2.3 ATTACK CHAIN ON MASKGCT

MaskGCT uses a very different architecture from that of F5-TTS and E2-TTS. It decouples the generation process into two main phases:

- **Text-to-Semantic (T2S):** It predicts discrete semantic tokens from phoneme sequences via a prefix-conditioned transformer structure.
- **Semantic-to-Acoustic (S2A):** It predicts multi-layer discrete acoustic codebook tokens conditioned on the semantic tokens via a generative transformer.

Both phases can be targeted by JASDA on the following two vulnerabilities.

Bidirectional attention mechanism. MaskGCT shifts from a causal attention mechanism to a bidirectional attention mechanism, which changes how the adversarial error propagates. In standard autoregressive systems, the causal mask forces errors to cascade strictly forward in time. In MaskGCT, the bidirectional mask removes this temporal constraint and transforms the vulnerability landscape in three unique ways.

(1) Unconstrained Global Jacobian Matrix: In a causal transformer, the attention weight matrix A (omitting indices $(l), m$) is strictly lower-triangular because a causal mask $\mathcal{M}_{\text{cau}}$ forces all future-facing elements to $-\infty$ before the softmax operation (see Eq. 13), which ensures:

$$\frac{\mathrm{d}A_{ij}}{\mathrm{d}B_{ik}} = 0, \quad \forall j > i \quad \text{or} \quad \forall k > i. \tag{37}$$

It restricts the influence of a perturbation to downstream positions ($i \geqslant k$). However, in the **bidirectional configuration** of MaskGCT, the attention matrix A is **fully dense**, which opens up the entire global Jacobian matrix. A localized adversarial perturbation injected by JASDA into the reference prompt tokens yields a fully connected differential field where **every row i and every column j are dynamically linked**:

$$\mathrm{d}A_{ij} = A_{ij} \left(\mathrm{d}B_{ij} - \sum_{k=1}^{T} A_{ik}\, \mathrm{d}B_{ik} \right), \quad \forall i, j. \tag{38}$$

Because there are no structural zeroes in this system, the perturbation spreads across the entire sequence space in a single forward pass, regardless of temporal ordering.

(2) Bidirectional Latent Representation Disruption: We formulate the joint perturbed latent representation $\tilde{z}$ as a block concatenation of the reference prompt embeddings $\tilde{z}_P \in \mathbb{R}^{T_P \times \mathfrak{d}_m}$ and the generation target tokens $z_G \in \mathbb{R}^{T_G \times \mathfrak{d}_m}$: $\tilde{z} := \begin{bmatrix} \tilde{z}_P \\ z_G \end{bmatrix}$. Because attention is bidirectional, the pre-activation attention matrix $\tilde{B}$ includes an unconstrained cross-modal block $\tilde{B}_{GP}$ that represents how the target positions attend back to the reference prompt:

$$\tilde{B}_{GP} = \frac{Q_G \tilde{K}_P^\top}{\sqrt{\mathfrak{d}_m}} = \frac{(z_G W^Q)(\tilde{z}_P W^K)^\top}{\sqrt{\mathfrak{d}_m}}. \tag{39}$$

When JASDA alters the reference audio to introduce an input embedding perturbation $\mathrm{d}z_P \approx \delta$, the direct differential impact on $\tilde{B}_{GP}$ is derived as:

$$\mathrm{d}B_{GP} = \frac{Q_G \mathrm{d}K_P^\top}{\sqrt{\mathfrak{d}_m}} \approx \frac{(z_G W^Q)\left(\delta W^K\right)^\top}{\sqrt{\mathfrak{d}_m}}. \tag{40}$$

In this bidirectional setting, the perturbation δ is linearly scaled by the target queries Q_G and consequently disrupts the entire generation space.

(3) Cyclic Vulnerability in Mask-predict Loops: While continuous architectures like F5-TTS and E2-TTS diffuse perturbations across ODE trajectories, MaskGCT executes a discrete and iterative Mask-predict generation routine. The intersection of unconstrained bidirectional attention and discrete token-unmasking mechanics exposes a severe cyclic vulnerability:

- **Forward Error Injection:** During decoding iteration t, the adversarial features embedded in the reference prompt distort the decoding confidence score vector $\mathcal{S}$. This forces the parallel decoder to incorrectly isolate and unmask a subset of highly corrupted target tokens at spatial coordinates defined by the mask index operator $\mathcal{M}_t$.

- **Reciprocal Backward Reflection:** In the subsequent iteration $(t+1)$, these newly unmasked and corrupted tokens are no longer obscured by the mask token embedding. Because the attention topology is bidirectional, these corrupted target elements are projected onto the active hidden states, which function as distorted Keys and Values for the global sequence.
- **Cyclic Amplification Loop:** In all succeeding iterations, these active and corrupted target coordinates project their distorted distributions non-causally, which impacts both preceding tokens ($i < \mathcal{M}_t$) and succeeding tokens ($i > \mathcal{M}_t$) simultaneously.

Decoding trajectory hijacking. Suppose the output layer maps $\tilde{z}$ to the final output logits as $\tilde{\mathscr{L}} \in \mathbb{R}^{T \times \mathfrak{d}_{\mathcal{V}}}$, where $\mathfrak{d}_{\mathcal{V}}$ is the codebook vocabulary size. In the MaskGCT implementation, the decoding confidence score vector $\tilde{\mathcal{S}}$ is obtained by:

$$\tilde{\mathcal{S}} := \mathrm{gather}(\mathrm{softmax}(\tilde{\mathscr{L}}), \mathbb{J}_{spl}), \tag{41}$$

where gather is the gathering operator of `PyTorch` with sample indices $\mathbb{J}_{spl}$. The differential of these confidence scores scales with the attention layer perturbation:

$$\mathrm{d}\mathcal{S} \propto \nabla_{\tilde{z}}\mathcal{S} \cdot \left(\mathrm{d}A \cdot \tilde{V} + \tilde{A} \cdot \mathrm{d}V \right), \tag{42}$$

where $\tilde{A}$ and $\tilde{V}$ denote the perturbed attention score matrix and the value matrix, respectively. MaskGCT updates its active decoding mask at step t using a rigid hard thresholding operation:

$$\tilde{\mathcal{M}}_t = \mathbb{J}_{tpk}(1 - \tilde{\mathcal{S}}, N_{tpk,t}), \tag{43}$$

where $\mathbb{J}_{tpk}$ (the top-k operator) picks up the $N_{tpk,t}$ tokens that MaskGCT is least confident about and marks them to be re-predicted or masked out at step t. Because the structured perturbation δ of JASDA induces a targeted shift in the confidence scores $\mathrm{d}\mathcal{S}$, and $\mathbb{J}_{tpk}$ contains discontinuities at boundary crossings where score values may swap order:

$$\exists \delta, i \quad \text{s.t.} \quad \mathrm{rank}\left(\mathcal{S}_i + \mathrm{d}\mathcal{S}_i\right) \neq \mathrm{rank}(\mathcal{S}_i), \tag{44}$$

it may alter the discrete set of indices output by $\tilde{\mathcal{M}}_t$. Once an incorrect subset of acoustic tokens is chosen for early unmasking, their corrupted values serve as fixed contextual keys/values in subsequent decoding iterations $(t+1), \ldots, T$. This may cause immediate compounding disruptions in acoustic generation.

A.2.4 Attack Chain on ZipVoice

ZipVoice adopts an optimized Transformer architecture designed for speech sequences called Zipformer (Yao et al., 2024) with a CFM decoder. As for inference, ZipVoice considers speech synthesis as a continuous normalizing flow problem by solving an ODE via an Euler Solver (Lipman et al., 2023). The vector field predictions are heavily conditioned on text conditions and speech prompt conditions, which can be integrated as $\mathcal{C}$ without loss of generality. These mechanisms can be targeted by JASDA on at least three vulnerabilities.

Error accumulation along the time dimension. Let $\tilde{z}_t$ be the perturbed acoustic feature trajectory at time step $t \in [0, 1]$. The time-dependent vector field is defined by the Zipformer parameter network $v_{zip}(\tilde{z}_t, t, \tilde{\mathcal{C}})$. The true generation trajectory is governed by the probability flow ODE, which can be numerically solved by the Euler Solver over $\mathfrak{N}$ discrete steps. Let $\Delta t = \frac{1}{\mathfrak{N}}$, then the discrete forward update from step t to $t + \Delta t$ is:

$$\tilde{z}_{t+\Delta t} = \tilde{z}_t + v_{zip}(\tilde{z}_t, t, \tilde{\mathcal{C}}) \cdot \Delta t. \tag{45}$$

It indicates that $\tilde{z}_{t+\Delta t}$ not only starts at the perturbed $\tilde{z}_t$, but also uses the perturbed Zipformer $v_{zip}(\tilde{z}_t, t, \tilde{\mathcal{C}})$.

Attention layer perturbation dynamics. In the `TTSZipformer` layers of ZipVoice, the interaction between the latent trajectory $\tilde{z}$ and the conditions $\tilde{\mathcal{C}}$ is **mediated** via attention blocks:

$$\tilde{Q} = \tilde{z}W^Q, \quad \tilde{K} = \tilde{\mathcal{C}}W^K, \quad \tilde{V} = \tilde{\mathcal{C}}W^V. \tag{46}$$

The corresponding attention matrix $\tilde{A}$ and the layer output $\tilde{\zeta}$ are:

$$\tilde{A} = \text{softmax}\left(\frac{\tilde{Q}\tilde{K}^\top}{\sqrt{\mathfrak{d}_m}}\right), \quad \tilde{\zeta} = \tilde{A}\tilde{V}. \tag{47}$$

Taking the total differential of $\tilde{\zeta}$ leads to:

$$\mathrm{d}\tilde{\zeta} = \mathrm{d}A \cdot \tilde{V} + \tilde{A} \cdot \mathrm{d}V, \tag{48}$$

Since the perturbation is injected into the reference speech, which belongs to the condition $\tilde{\mathcal{C}}$, we can denote $\tilde{\mathcal{C}} = \mathcal{C} + \delta$ and $\mathrm{d}\mathcal{C} \approx \delta$ without loss of generality. Then taking differentials of the latter two terms in (46) leads to:

$$\mathrm{d}K \approx \delta W^K, \quad \mathrm{d}V \approx \delta W^V. \tag{49}$$

On the other hand, since $\mathrm{d}K \propto \delta$, direct calculation yields

$$\mathrm{d}A_i = \left[\frac{\tilde{Q}(\mathrm{d}K)^\top}{\sqrt{\mathfrak{d}_m}}\right]_i (\mathrm{diag}(\tilde{A}_i) - \tilde{A}_i^\top \tilde{A}_i) \propto \delta, \tag{50}$$

where the subscript i denotes the i-th row of a matrix. Gathering all i, we also have

$$\mathrm{d}A \propto \delta. \tag{51}$$

Inserting $\mathrm{d}V \propto \delta$ and $\mathrm{d}A \propto \delta$ into (48), we have

$$\mathrm{d}\tilde{\zeta} \propto \nabla_{\mathcal{C}}\zeta \cdot \delta. \tag{52}$$

Hence the localized structural perturbation at $\tilde{\zeta}$ can be represented as a linear function of δ.

Accumulated error propagation through the Euler solver. Because the vector field network v_{zip} takes ζ as an intermediate representation, the perturbation propagates directly into the velocity field computation. Let $\tilde{v}_{zip}$ be the perturbed vector field:

$$\tilde{v}_{zip}(\tilde{z}_t, t, \mathcal{C} + \delta) = v_{zip}(\tilde{z}_t, t, \mathcal{C}) + \nabla_\zeta v_{zip}(\tilde{z}_t, t, \mathcal{C}) \cdot \mathrm{d}\zeta + \mathcal{O}(\|\delta\|_2^2). \tag{53}$$

It further worsens the forward update of (45):

$$\tilde{z}_{t+\Delta t} = \tilde{z}_t + \tilde{v}_{zip}(\tilde{z}_t, t, \mathcal{C} + \delta) \cdot \Delta t. \tag{54}$$

Let $e_t = \tilde{z}_t - z_t$ be the accumulated trajectory error at step t. Subtracting the unperturbed equation from the perturbed equation gives the error recurrence relation:

$$e_{t+\Delta t} = e_t + [\tilde{v}_{zip}(\tilde{z}_t, t, \mathcal{C} + \delta) - v_{zip}(z_t, t, \mathcal{C})] \cdot \Delta t. \tag{55}$$

Applying a multivariate first-order Taylor expansion to $v_{zip}(\tilde{z}_t, t, \mathcal{C})$ around the clean state z_t yields:

$$v_{zip}(\tilde{z}_t, t, \mathcal{C}) \approx v_{zip}(z_t, t, \mathcal{C}) + \nabla_z v_{zip}(z_t, t, \mathcal{C}) \cdot e_t. \tag{56}$$

Substituting (56) and (48) into (55) yields

$$\begin{aligned} e_{t+\Delta t} &= e_t + [\nabla_z v_{zip}(z_t, t, \mathcal{C}) \cdot e_t + \nabla_\zeta v_{zip}(\tilde{z}_t, t, \mathcal{C}) \cdot \mathrm{d}\zeta] \cdot \Delta t \\ &= (\mathbb{I} + \nabla_z v_{zip}(z_t, t, \mathcal{C}) \cdot \Delta t) \cdot e_t + [\nabla_\zeta v_{zip}(\tilde{z}_t, t, \mathcal{C}) \cdot (\mathrm{d}A \cdot V + A \cdot \mathrm{d}V)] \cdot \Delta t, \end{aligned} \tag{57}$$

which indicates the error accumulation path.

A.3 Eight Defense Methods

A.3.1 Adversarial Training (AT)

The AT defense aims to enforce direct feature similarity between the transformer representation of **adversarial target** sample $\tilde{h}_{tgt}$ and **clean target** sample h_{tgt}. To maintain clean synthesis quality, a regularizing term comparing **clean source** latent representation h_{src} against clean target latent representation h_{tgt} is weighted by a hyperparameter β. The corresponding loss function for AT is:

$$\mathcal{L}_{\text{AT}} = \|\tilde{h}_{tgt} - h_{tgt}\|_2^2 + \beta \cdot \|h_{src} - h_{tgt}\|_2^2. \tag{58}$$

A.3.2 KULLBACK-LEIBLER (KL) DIVERGENCE REGULARIZATION

KL regularization smooths feature activation shifts caused by adversarial perturbations by aligning soft probability distributions. The hidden states are processed by the layer normalization operator $\mathcal{T}_{LN}$ and scaled by temperature $\mathfrak{T}_{KL}$. Then the clean and adversarial target distributions can be formulated as:

$$p_{cln} = \mathrm{softmax}\left(\frac{\mathcal{T}_{LN}(h_{tgt})}{\mathfrak{T}_{KL}}\right), \quad p_{adv} = \mathrm{softmax}\left(\frac{\mathcal{T}_{LN}(\tilde{h}_{tgt})}{\mathfrak{T}_{KL}}\right). \tag{59}$$

The full regularized KL loss combines temperature-scaled KL divergence with feature matching losses weighted by a hyperparameter α:

$$\begin{aligned}
&\mathcal{L}_{div} = \mathfrak{T}_{KL}^2 \cdot \mathcal{D}_{KL}\left(p_{adv} \parallel p_{cln}\right), \\
&\mathcal{L}_{KL} = \mathcal{L}_{\text{div}} + \alpha \cdot \|\mathcal{T}_{LN}(\tilde{h}_{tgt}) - \mathcal{T}_{LN}(h_{tgt})\|_2^2 + \alpha \cdot \|\mathcal{T}_{LN}(h_{src}) - \mathcal{T}_{LN}(h_{tgt})\|_2^2,
\end{aligned} \tag{60}$$

where $\mathcal{D}_{KL}$ denotes the KL divergence.

A.3.3 GRIFFIN-LIM (GL) DEFENSE

Recall that $\tilde{x}_{ref}$ is the perturbed reference audio prompt by JASDA. The GL defense aims to transform $\tilde{x}_{ref}$ back into a valid physical audio manifold by reconstructing a consistent phase from the Short-Time Fourier Transform (STFT) magnitude. The discrete STFT operator $\mathscr{F}$ maps $\tilde{x}_{ref}$ to a complex time-frequency representation $\tilde{\mathcal{X}} \in \mathbb{C}^{M_F \times K_F}$, where M_F and K_F denote the numbers of frames and frequency bins, respectively.

$$\tilde{\mathcal{X}}[m,k] = \mathscr{F}(\tilde{x}_{ref})[m,k] = \sum_{n=0}^{N_w-1} \tilde{x}_{\text{ref}}[m \cdot \mathcal{H} + n]\, w[n]\, e^{-\mathrm{i}\frac{2\pi}{N_F}kn}, \tag{61}$$

where N_F is the frame size, $\mathcal{H}$ is the hop length, N_w is the window size, $w[n]$ is the Hann window function, i denotes the imaginary unit, $m \in \{0, \ldots, M_F - 1\}$ is the frame index, and $k \in \{0, \ldots, K_F - 1\}$ is the frequency bin index ($K_F = \frac{N_F}{2} + 1$).

Next, the complex spectrum $\tilde{\mathcal{X}}[m,k]$ is decomposed into magnitude $\tilde{\Xi}[m,k]$ and phase $\tilde{\Psi}[m,k]$:

$$\tilde{\Xi}[m,k] := |\tilde{\mathcal{X}}[m,k]|, \quad \tilde{\Psi}[m,k] := \angle\tilde{\mathcal{X}}[m,k]. \tag{62}$$

The GL algorithm discards or modifies the adversarial phase component $\tilde{\Psi}[m,k]$. The phase estimation sequence initializes with a random or baseline phase $\phi^{(0)}[m,k] \sim \mathcal{U}(-\pi, \pi)$, where $\mathcal{U}$ denotes the uniform distribution. At the o-th defense iteration, the GL algorithm reconstructs the estimated complex STFT matrix via:

$$\tilde{\mathcal{X}}^{(o)}[m,k] = \tilde{S}[m,k] \cdot \exp\left(\mathrm{i}\phi^{(o)}[m,k]\right), \quad \forall m,k. \tag{63}$$

Then the inverse STFT (iSTFT) is implemented as the synthesis operator:

$$\hat{x}^{(o)} = \mathscr{F}^{-1}\left(\tilde{\mathcal{X}}^{(o)}\right). \tag{64}$$

The phase of the synthesized signal is updated by:

$$\phi^{(o+1)}[m,k] = \angle\mathscr{F}\left(\hat{x}^{(o)}\right)[m,k]. \tag{65}$$

After a sufficient number O of defense iterations, the recovered reference prompt is $x_{rcv} := \hat{x}^{(O)}(1 : T)$.

A.3.4 LOW-PASS FILTERING (LPF) DEFENSE

LPF suppresses high-frequency adversarial perturbations under the assumption that human speech semantics and core identity features are concentrated in lower spectral bands, while residual adversarial perturbation energy δ is contained in high-frequency bands. Given the perturbed reference

audio signal $\tilde{x}_{ref}$ sampled at frequency $\mathfrak{f}_s$, the defense applies an N_B-th order continuous-time analog Butterworth LPF (Butterworth et al., 1930) characterized by a cutoff frequency $\mathfrak{f}_c$. The rational z-domain transfer function $H(z)$ is defined as:

$$H(z) := \frac{\sum_{n=0}^{N_B} b_n z^{-n}}{1 + \sum_{n=1}^{N_B} a_n z^{-n}}, \tag{66}$$

where $\{b_n\}_{n=0}^{N_B}$ and $\{a_n\}_{n=1}^{N_B}$ are the filter coefficients computed on $\mathfrak{f}_s$ and $\mathfrak{f}_c$.

To avoid group delay and non-linear phase distortion across speech formants, the defense employs a zero-phase forward-backward filtering procedure. Denote $\mathscr{F}$ and $\mathscr{F}^{-1}$ as the discrete-time Fourier transform (DTFT) and its inverse, respectively. The forward pass filters the perturbed signal $\tilde{x}_{ref}[t]$ to produce an intermediate signal $\tilde{\mathfrak{x}}_1[t]$, then reverses its time to yield $\tilde{\mathfrak{x}}_2[t]$:

$$\tilde{\mathfrak{x}}_1[t] = \mathscr{F}^{-1}\left\{H(\mathrm{e}^{\mathrm{i}\omega}) \cdot \mathscr{F}\{\tilde{x}_{ref}[t]\}\right\}, \quad \tilde{\mathfrak{x}}_2[t] = \tilde{\mathfrak{x}}_1[-t], \quad \forall t. \tag{67}$$

The backward pass further filters $\tilde{\mathfrak{x}}_2[t]$, then reverses its time back to yield the recovered signal x_{rcv}:

$$\tilde{\mathfrak{x}}_3[t] = \mathscr{F}^{-1}\left\{H(\mathrm{e}^{\mathrm{i}\omega}) \cdot \mathscr{F}\{\tilde{\mathfrak{x}}_2[t]\}\right\}, \quad x_{rcv}[t] = \tilde{\mathfrak{x}}_3[-t], \quad \forall t. \tag{68}$$

A.3.5 μ-LAW DEFENSE

Given a perturbed normalized waveform $\tilde{x}(t) \in [-1, 1]$, it is compressed by logarithmic companding to redistribute energy across dynamic levels. For a companding parameter $\mu > 0$, the compression mapping $\mathcal{T}_{cps} : [-1, 1] \to [-1, 1]$ is defined as:

$$\mathcal{T}_{cps}(\tilde{x}(t)) = \mathrm{sign}(\tilde{x}(t)) \cdot \frac{\log(1 + \mu|\tilde{x}(t)|)}{\log(1 + \mu)}. \tag{69}$$

Then $\mathcal{T}_{cps}(\tilde{x}(t))$ is quantized into $(\mu + 1)$ discrete levels across the interval:

$$\mathcal{T}^q_{cps}(\tilde{x}(t)) = \frac{2}{\mu} \cdot \mathrm{round}\left(\frac{\mu(\mathcal{T}_{cps}(\tilde{x}(t)) + 1)}{2}\right) - 1. \tag{70}$$

The inverse mapping $\mathcal{T}^{-1}_{cps}$ recovers the audio signal x_{rcv} from quantized space back into the linear amplitude domain:

$$x_{rcv}(t) = \mathcal{T}^{-1}_{cps}(\mathcal{T}^q_{cps}(\tilde{x}(t))) = \mathrm{sign}(\mathcal{T}^q_{cps}(\tilde{x}(t))) \cdot \frac{1}{\mu}\left((1 + \mu)^{|\mathcal{T}^q_{cps}(\tilde{x}(t))|} - 1\right). \tag{71}$$

A.3.6 RANDOMIZED SMOOTHING (RS) DEFENSE

RS aims to purify the perturbed reference waveform $\tilde{x}_{ref} = x_{ref} + \delta$ by convolving the input space with an isotropic Gaussian noise distribution $\mathcal{N}(0, \sigma^2_{rs}\mathbb{I}_{\mathfrak{d}_x})$:

$$x_{rs} = \mathrm{proj}_{[-1,1]^{\mathfrak{d}_x}}\left(\tilde{x}_{ref} + \xi\right) = \mathrm{proj}_{[-1,1]^{\mathfrak{d}_x}}\left(x_{ref} + \delta + \xi\right), \quad \xi \sim \mathcal{N}\left(0, \sigma^2_{rs}\mathbb{I}_{\mathfrak{d}_x}\right), \tag{72}$$

where $\sigma_{rs} > 0$ is the standard deviation (STD) for RS, $\mathfrak{d}_x$ is the dimensionality of x_{ref}, and $\mathrm{proj}_{[-1,1]^{\mathfrak{d}_x}}$ clamps the signal within valid normalized audio bounds.

A.3.7 TOTAL VARIATION (TV) REGULARIZATION DEFENSE

TV regularization intends to remove high-frequency adversarial perturbation noise from an audio signal $\tilde{x}_{ref}$ while preserving sharp signal transitions. The objective function is formulated as:

$$x^* = \arg\min_{x \in \mathbb{R}^T} \frac{1}{2}\|x - \tilde{x}_{ref}\|_2^2 + \lambda \mathcal{T}_{TV}(x), \quad \mathcal{T}_{TV} := \sum_{t=1}^{T-1} |x(t+1) - x(t)|, \tag{73}$$

where $\mathcal{T}_{TV}$ denotes the TV operator. This model can be solved via the primal-dual algorithm. We define the linear forward difference (gradient) operator $\nabla : \mathbb{R}^T \to \mathbb{R}^T$ as:

$$(\nabla x)_t = \begin{cases} x_{t+1} - x_t & \text{if } 1 \le t < T \\ 0 & \text{if } t = T \end{cases}. \tag{74}$$

The adjoint operator $\nabla^\top : \mathbb{R}^T \to \mathbb{R}^T$ satisfies $\langle \nabla x, u \rangle = \langle x, \nabla^\top u \rangle$ for standard inner products, which leads to:

$$(\nabla^\top u)_t = \begin{cases} -u_1 & \text{if } t = 1 \\ u_{t-1} - u_t & \text{if } 1 < t < T \\ u_{T-1} & \text{if } t = T \end{cases}. \tag{75}$$

Chambolle's projected gradient algorithm (Chambolle, 2004) iteratively solves for the dual variable $u^{(k)}$ with step-size parameter ς:

$$u_t^{(k+1)} = \frac{u_t^{(k)} + \varsigma \left[\nabla\left(-\nabla^\top u^{(k)} + \frac{\tilde{x}_{ref}}{\lambda}\right)\right]_t}{1 + \varsigma \left|\left[\nabla\left(-\nabla^\top u^{(k)} + \frac{\tilde{x}_{ref}}{\lambda}\right)\right]_t\right|}, \quad k = 1, 2, 3, \cdots \tag{76}$$

This algorithm outputs a solution of dual variable u^*. Then the recovered audio signal (as the primal variable) can be calculated by:

$$x_{rcv} = \tilde{x}_{ref} - \lambda \nabla^\top u^*. \tag{77}$$

A.3.8 Score-based Speech Enhancement (SSE)

SSE employs continuous-time stochastic differential equations (SDEs) to purify perturbed speech by modeling the score (the gradient of the log-probability density) of the speech feature distribution. Given the perturbed reference audio prompt $\tilde{x}_{ref} \in \mathbb{R}^T$, the defense scheme first converts the raw time-domain signal into a complex STFT representation:

$$\tilde{\mathcal{X}} = \mathscr{F}(\tilde{x}_{ref}) \in \mathbb{C}^{M_F \times K_F}. \tag{78}$$

The backward denoising process progressively transforms $\tilde{\mathcal{X}}$ toward clean speech features $\mathcal{X}_0$ from $t = 1$ to $t = 0$ via an inverse Ornstein-Uhlenbeck-type SDE:

$$\mathrm{d}\mathcal{X}_t = \left[-\gamma(\tilde{\mathcal{X}} - \mathcal{X}_t) + g(t)^2 \nabla_{\mathcal{X}} \log p_t(\mathcal{X}_t \mid \tilde{\mathcal{X}})\right] dt + g(t)\, \mathrm{d}\bar{\mathcal{W}}_t, \tag{79}$$

where $\mathcal{X}_t \in \mathbb{C}^{M_F \times K_F}$ is the latent state at time t, $\gamma > 0$ controls the mean-reverting drift, $g(t)$ is the time-dependent diffusion coefficient, and $\bar{\mathcal{W}}_t$ is a complex backward Wiener process. The conditional score vector field $\nabla_{\mathcal{X}} \log p_t(\mathcal{X}_t \mid \tilde{\mathcal{X}})$ is approximated by a pretrained deep neural score network $s_\theta(\mathcal{X}_t, \tilde{\mathcal{X}}, t)$:

$$\nabla_{\mathcal{X}} \log p_t(\mathcal{X}_t \mid \tilde{\mathcal{X}}) \approx s_\theta(\mathcal{X}_t, \tilde{\mathcal{X}}, t). \tag{80}$$

After the backward process, the recovered audio signal can be obtained by the iSTFT:

$$x_{rcv} = \mathscr{F}^{-1}(\mathcal{X}_0). \tag{81}$$

Remark: Defense parameters (e.g., regularization strength, filter cutoff frequency, noise scale) are tuned based on preliminary experiments for each target model to achieve optimal trade-offs between defense effectiveness and audio quality preservation. AT and KL require white-box access for defensive fine-tuning and are evaluated on all five target models. The attacker does not require access to the target models.

A.4 Experimental Settings

A.4.1 Datasets and Test Samples

LibriSpeech-PC (Meister et al., 2023): It is a speech recognition corpus designed to train and evaluate end-to-end automatic speech recognition models on joint speech recognition, punctuation, and capitalization prediction. We randomly sample 100 sentences from the test-clean split to establish the baseline performance. This subset serves as the primary benchmark for comparing attack effectiveness across different target models.

Emilia (He et al., 2024): It is a large-scale and multilingual speech dataset designed for training natural and spontaneous speech generation models, sourcing from *in-the-wild* media such as podcasts

and online videos. It captures diverse and highly spontaneous human speaking styles rather than rigid and formal audiobook readings. We extract 100 English text samples from this dataset.

Common Voice (Ardila et al., 2020): It is a massively-multilingual collection of transcribed speech designed for speech technology research. The corpus is built through community crowdsourcing, where volunteers read and record displayed text prompts via a web interface or app, while other contributors validate the recordings using a peer-voting mechanism. We randomly sample 100 English utterances by following the same protocol to validate cross-dataset robustness.

For each dataset, all five target models synthesize speech with the *same* 100 text prompts to ensure fair comparison.

A.4.2 Attack Configuration

A single universal adversarial prompt is applied to all target models to simulate a realistic black-box threat scenario. The clean reference prompt is a 4.455-second English utterance sampled at 16kHz in floating-point format, with the transcript: *where I am in the warm room with all the splendor and magnificence*. The adversarial reference prompt is generated at $24kHz$ with PCM_16 encoding. Both audio prompts have the same duration.

The adversarial perturbation is constrained under the ℓ_∞ norm with budget $\epsilon = 0.01$ when audio waveforms are normalized to $[-1, 1]$. This budget corresponds to an input SNR of approximately 19dB. The attack is executed with $\mathcal{K} = 200$ iterations and step size $\eta = 0.002$.

A key advantage of JASDA is that the adversarial prompt is crafted *for only once* and then input into all target models without any model-specific adaptation or query feedback, which makes it a true black-box attack under zero-interaction assumption.

A.4.3 Evaluation Metrics

Speaker Similarity (SIM). The SIM metric is defined as the cosine similarity between the two embeddings e_{gen} and e_{ref} of the generated and the *clean* reference audios.

$$e_{gen}{=}f_{spk}(x_{gen}), e_{ref}{=}f_{spk}(x_{ref}), \mathrm{SIM}(x_{gen}, x_{ref}){:=}\cos(e_{gen}, e_{ref}){=}\frac{\langle e_{gen}, e_{ref}\rangle}{\|e_{gen}\|_2\|e_{ref}\|_2}, \quad (82)$$

where x_{gen} and x_{ref} denote the generated and the reference audios, respectively. f_{spk} denotes the SpeechBrain ECAPA-TDNN model (Dawalatabad et al., 2021) pretrained on VoxCeleb (Nagrani et al., 2020), which serves as the embedding extractor. SIM ranges in $[-1, 1]$, with higher values indicating better preservation of speaker identity.

Relative SIM Drop (RSD). This metric quantifies the severity of speaker identity degradation:

$$\mathrm{RSD}(attack){=}\frac{\mathrm{SIM}_{clean} - \mathrm{SIM}_{attack}}{\mathrm{SIM}_{clean}}{\times}100\%, \mathrm{RSD}(defense){=}\frac{\mathrm{SIM}_{clean} - \mathrm{SIM}_{defense}}{\mathrm{SIM}_{clean}}{\times}100\%. \quad (83)$$

It normalizes the absolute SIM decrease by the clean baseline to enable fair comparison across different target models with different intrinsic SIM levels. This is the primary metric that assesses the **destructive capacity** of JASDA.

A.4.4 Implementation Details

Hardware and Software: All experiments are conducted on an NVIDIA Pro6000 GPU with CUDA 13.0. We use PyTorch 2.11.0 for model inference and adversarial attack implementation. The ECAPA-TDNN speaker verification model is loaded from the SpeechBrain toolkit (Ravanelli et al., 2024).

Audio Processing: Target models operate at their native sampling rates ($16kHz$ or $24kHz$). The ZipVoice inference pipeline automatically resamples prompts to $24kHz$ according to model configuration. All generated outputs are resampled to a common rate for SIM evaluation to ensure consistency.

Baseline Anchors: Each dataset-model combination maintains a clean baseline anchor. For Emilia experiments, we use same-machine clean baselines to avoid environment-induced variability. RSD is computed using the corresponding clean anchor for each experimental configuration.

A.5 Additional Experiments

A.5.1 Main Results on Emilia

Table A1 evaluates the attack performance of JASDA across five SOTA NAR TTS models on the Emilia dataset, alongside the efficacy of eight defense strategies.

Attack effectiveness across target models. In the undefended setting, the JASDA attack significantly degrades speaker cloning fidelity across all target architectures relative to the clean baseline.
• *High susceptibility in flow-matching models:* Flow-matching and diffusion-based NAR models expose the highest vulnerability to JASDA. Specifically, E2-TTS suffers the largest performance deterioration, with SIM dropping from 0.7770 to 0.2953 (ΔSIM $= -0.4817$, a 61.99% RSD). Similarly, F5-TTS drops from 0.7240 to 0.2932 (ΔSIM $= -0.4308$, a 59.50% RSD).
• *Moderate vulnerability in hybrid models:* LEMAS-TTS and ZipVoice also experience moderate drops under attack, declining by 43.96% (SIM drops from 0.7705 to 0.4318) and 39.54% (SIM drops from 0.7377 to 0.4460), respectively.
• *Inherent resilience in codec-based generation:* MaskGCT shows the strongest intrinsic robustness against JASDA, and maintains a SIM of 0.5290 under attack compared to its 0.6860 clean baseline. This represents the lowest RSD (22.89%) among all target models.

Defense efficacy analysis. Evaluating the defensive countermeasures against JASDA reveals substantial variations in defending capability.
• *Filtering and enhancement defenses (LPF, SSE):* LPF emerges as the most robust baseline defense and yields positive ΔSIM gains across all five models. Specifically, LPF achieves the highest SIM recovery on E2-TTS (ΔSIM $= +0.1757$, recovering SIM to 0.4710) and F5-TTS (ΔSIM $= +0.1045$, recovering SIM to 0.3977). SSE similarly offers effective recovery on LEMAS-TTS (ΔSIM $= +0.0814$) and E2-TTS (ΔSIM $= +0.0780$), but shows an anomalous performance drop on MaskGCT (ΔSIM $= -0.1305$, RSD increasing to 41.91%). However, **neither defense can recover SIM to the clean level**.
• *Transform-based defenses (TV, GL):* TV and GL provide modest improvements against JASDA across all models, which achieve positive ΔSIM gains ranging from $+0.0036$ to $+0.0705$.
• *Ineffective defenses (μ-law, RS):* μ-law and RS provide little or no defensive benefit against JASDA perturbations and produce near zero change in SIM score across all target models (ΔSIM ≈ 0.000). Moreover, RS even worsens speaker similarity slightly on LEMAS-TTS.
• *Counterproductive defenses (AT, KL):* AT and KL fail to defend against JASDA and frequently exacerbate degradation. On LEMAS-TTS, AT further reduces SIM to 0.3332 (ΔSIM $= -0.0986$), and on E2-TTS, AT leads to the worst overall performance across the benchmark (0.2581 SIM, 66.78% RSD).

A.5.2 Main Results on Common Voice

Table A2 presents the experimental evaluation of the JASDA attack across five SOTA NAR TTS models under clean conditions, undefended attack settings, and eight defense countermeasures on the Common Voice dataset.

Attack efficacy under no defense. The experimental results show that JASDA substantially degrades SIM across all five target models. In the absence of defense mechanisms, JASDA incurs severe performance degradation and achieves an average RSD exceeding 44%. F5-TTS and E2-TTS expose the highest vulnerability to JASDA and suffer RSDs of 60.18% (SIM falls from 0.7115 to 0.2833, ΔSIM $= -0.4282$) and 54.31% (SIM falls from 0.7626 to 0.3484, ΔSIM $= -0.4142$), respectively. ZipVoice and LEMAS-TTS also undergo considerable degradation, with RSDs of 40.03% and 42.90%. While MaskGCT shows greater inherent resilience, it still incurs a significant 25.61% RSD in SIM score (0.6803 to 0.5061), which indicates the broad effectiveness of JASDA across diverse model structures.

Defense countermeasure analysis. Evaluating the eight defense measures reveals differences in defending efficacy and stability against JASDA:
• *Moderate-efficacy defense:* LPF acts as the most robust defense across all five target architectures. It provides the highest absolute SIM recovery across the board, raising SIM by $+0.1267$ on E2-TTS,

Table A1: SIM scores for five target models under the JASDA attack and eight defense measures on the Emilia dataset. Each defense cell reports SIM, paired ΔSIM, and RSD.

Defense	ZipVoice	LEMAS-TTS	MaskGCT	F5-TTS	E2-TTS
Clean	SIM **0.7377**	SIM **0.7705**	SIM **0.6860**	SIM **0.7240**	SIM **0.7770**
No Defense	SIM **0.4460** $\Delta\text{SIM}_{\text{atk}}$ -0.2917 RSD 39.54%	SIM **0.4318** $\Delta\text{SIM}_{\text{atk}}$ -0.3387 RSD 43.96%	SIM **0.5290** $\Delta\text{SIM}_{\text{atk}}$ -0.1570 RSD 22.89%	SIM **0.2932** $\Delta\text{SIM}_{\text{atk}}$ -0.4308 RSD 59.50%	SIM **0.2953** $\Delta\text{SIM}_{\text{atk}}$ -0.4817 RSD 61.99%
AT	SIM **0.4520** $\Delta\text{SIM}_{\text{def}}$ +0.0060 RSD 38.73%	SIM **0.3332** $\Delta\text{SIM}_{\text{def}}$ -0.0986 RSD 56.76%	SIM **0.5337** $\Delta\text{SIM}_{\text{def}}$ +0.0047 RSD 22.20%	SIM **0.2908** $\Delta\text{SIM}_{\text{def}}$ -0.0024 RSD 59.83%	SIM **0.2581** $\Delta\text{SIM}_{\text{def}}$ -0.0372 RSD 66.78%
KL	SIM **0.4233** $\Delta\text{SIM}_{\text{def}}$ -0.0227 RSD 42.62%	SIM **0.3466** $\Delta\text{SIM}_{\text{def}}$ -0.0852 RSD 55.02%	SIM **0.5480** $\Delta\text{SIM}_{\text{def}}$ +0.0190 RSD 20.12%	SIM **0.2789** $\Delta\text{SIM}_{\text{def}}$ -0.0143 RSD 61.48%	SIM **0.3677** $\Delta\text{SIM}_{\text{def}}$ +0.0724 RSD 52.68%
TV	SIM **0.4644** $\Delta\text{SIM}_{\text{def}}$ +0.0184 RSD 37.05%	SIM **0.4825** $\Delta\text{SIM}_{\text{def}}$ +0.0507 RSD 37.38%	SIM **0.5427** $\Delta\text{SIM}_{\text{def}}$ +0.0137 RSD 20.89%	SIM **0.3637** $\Delta\text{SIM}_{\text{def}}$ +0.0705 RSD 49.77%	SIM **0.3272** $\Delta\text{SIM}_{\text{def}}$ +0.0319 RSD 57.89%
GL	SIM **0.4725** $\Delta\text{SIM}_{\text{def}}$ +0.0265 RSD 35.95%	SIM **0.4702** $\Delta\text{SIM}_{\text{def}}$ +0.0384 RSD 38.97%	SIM **0.5326** $\Delta\text{SIM}_{\text{def}}$ +0.0036 RSD 22.36%	SIM **0.3488** $\Delta\text{SIM}_{\text{def}}$ +0.0556 RSD 51.82%	SIM **0.3431** $\Delta\text{SIM}_{\text{def}}$ +0.0478 RSD 55.84%
μ-law	SIM **0.4451** $\Delta\text{SIM}_{\text{def}}$ -0.0009 RSD 39.66%	SIM **0.4312** $\Delta\text{SIM}_{\text{def}}$ -0.0006 RSD 44.04%	SIM **0.5222** $\Delta\text{SIM}_{\text{def}}$ -0.0068 RSD 23.88%	SIM **0.2923** $\Delta\text{SIM}_{\text{def}}$ -0.0009 RSD 59.63%	SIM **0.2953** $\Delta\text{SIM}_{\text{def}}$ 0.0000 RSD 61.99%
LPF	SIM **0.4554** $\Delta\text{SIM}_{\text{def}}$ +0.0094 RSD 38.27%	SIM **0.5048** $\Delta\text{SIM}_{\text{def}}$ +0.0730 RSD 34.48%	SIM **0.5672** $\Delta\text{SIM}_{\text{def}}$ +0.0382 RSD 17.32%	SIM **0.3977** $\Delta\text{SIM}_{\text{def}}$ +0.1045 RSD 45.07%	SIM **0.4710** $\Delta\text{SIM}_{\text{def}}$ +0.1757 RSD 39.38%
RS	SIM **0.4455** $\Delta\text{SIM}_{\text{def}}$ -0.0005 RSD 39.61%	SIM **0.4211** $\Delta\text{SIM}_{\text{def}}$ -0.0107 RSD 45.35%	SIM **0.5294** $\Delta\text{SIM}_{\text{def}}$ +0.0004 RSD 22.83%	SIM **0.2887** $\Delta\text{SIM}_{\text{def}}$ -0.0045 RSD 60.12%	SIM **0.2948** $\Delta\text{SIM}_{\text{def}}$ -0.0005 RSD 62.06%
SSE	SIM **0.4650** $\Delta\text{SIM}_{\text{def}}$ +0.0190 RSD 36.97%	SIM **0.5132** $\Delta\text{SIM}_{\text{def}}$ +0.0814 RSD 33.39%	SIM **0.3985** $\Delta\text{SIM}_{\text{def}}$ -0.1305 RSD 41.91%	SIM **0.3636** $\Delta\text{SIM}_{\text{def}}$ +0.0704 RSD 49.78%	SIM **0.3733** $\Delta\text{SIM}_{\text{def}}$ +0.0780 RSD 51.96%

$+0.1259$ on F5-TTS, $+0.0712$ on ZipVoice, $+0.0692$ on LEMAS-TTS, and $+0.0469$ on MaskGCT. This reduces the RSD of MaskGCT to 18.71%.

• *Model-dependent defenses:* SSE obtains considerable SIM gains on F5-TTS ($+0.1220$), ZipVoice ($+0.0666$), LEMAS-TTS ($+0.0619$), and E2-TTS ($+0.0608$). However, it fails sharply on MaskGCT, where SIM drops by 0.1914 (worsening the RSD to 53.74%), which indicates severe architecture-dependent friction.

• *Ineffective countermeasures:* Standard audio transformations such as μ-law and RS yield negligible defense gains ($\Delta\text{SIM} < 0.008$ across all models). Furthermore, conventional AT and KL show unpredictable behavior and exacerbate SIM degradation on LEMAS-TTS ($\Delta\text{SIM} = -0.0690$ for AT) and E2-TTS ($\Delta\text{SIM} = -0.0328$ for AT).

In summary, JASDA poses a severe threat to zero-shot voice cloning models. While LPF shows partial resilience, heuristic signal transformations fail to defend against JASDA perturbations.

A.5.3 Ablation Analysis

We further conduct ablation analysis with LEMAS-TTS on LibriSpeech-PC to examine whether the effectiveness of the JASDA perturbation depends on its optimized direction and temporal alignment. Let x_c be the clean waveform and $\delta_\star$ be the JASDA perturbation. Each ablated version uses the same additive RS waveform ξ as in Section A.3.6. The following ablated versions are evaluated:

$$\begin{aligned}
&\text{C1} \quad x_c + \xi && \text{(Clean + RS)},\\
&\text{M1} \quad x_c + \delta_\star + \xi && \text{(Ours)},\\
&\text{M2} \quad x_c + u + \xi && \text{(Random-}\ell_\infty\text{)},\\
&\text{M3} \quad x_c + |\delta_\star| \odot r + \xi && \text{(Sign-randomized Direction)},\\
&\text{M4} \quad x_c + \text{roll}(\delta_\star, 24000) + \xi && \text{(Temporal Shift)},
\end{aligned} \tag{84}$$

Table A2: SIM scores for five target models under the JASDA attack and eight defense measures on the Common Voice dataset. Each defense cell reports SIM, paired ΔSIM, and RSD.

Defense	ZipVoice	LEMAS-TTS	MaskGCT	F5-TTS	E2-TTS
Clean	SIM **0.7630**	SIM **0.7709**	SIM **0.6803**	SIM **0.7115**	SIM **0.7626**
No Defense	SIM **0.4576** $\Delta\mathrm{SIM}_{\mathrm{atk}}$ -0.3054 RSD 40.03%	SIM **0.4402** $\Delta\mathrm{SIM}_{\mathrm{atk}}$ -0.3307 RSD 42.90%	SIM **0.5061** $\Delta\mathrm{SIM}_{\mathrm{atk}}$ -0.1742 RSD 25.61%	SIM **0.2833** $\Delta\mathrm{SIM}_{\mathrm{atk}}$ -0.4282 RSD 60.18%	SIM **0.3484** $\Delta\mathrm{SIM}_{\mathrm{atk}}$ -0.4142 RSD 54.31%
AT	SIM **0.4795** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0219 RSD 37.16%	SIM **0.3712** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0690 RSD 51.85%	SIM **0.5207** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0146 RSD 23.46%	SIM **0.2964** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0131 RSD 58.34%	SIM **0.3156** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0328 RSD 58.62%
KL	SIM **0.4575** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0001 RSD 40.04%	SIM **0.3892** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0510 RSD 49.51%	SIM **0.5210** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0149 RSD 23.42%	SIM **0.2803** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0030 RSD 60.60%	SIM **0.4162** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0678 RSD 45.42%
TV	SIM **0.4898** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0322 RSD 35.81%	SIM **0.4969** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0567 RSD 35.54%	SIM **0.5337** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0276 RSD 21.55%	SIM **0.3679** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0846 RSD 48.29%	SIM **0.3251** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0233 RSD 57.37%
GL	SIM **0.4850** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0274 RSD 36.44%	SIM **0.4825** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0423 RSD 37.41%	SIM **0.5095** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0034 RSD 25.11%	SIM **0.3035** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0202 RSD 57.34%	SIM **0.3347** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0137 RSD 56.11%
μ-law	SIM **0.4571** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0005 RSD 40.09%	SIM **0.4407** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0005 RSD 42.83%	SIM **0.5085** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0024 RSD 25.25%	SIM **0.2832** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0001 RSD 60.20%	SIM **0.3463** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0021 RSD 54.59%
LPF	SIM **0.5288** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0712 RSD 30.69%	SIM **0.5094** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0692 RSD 33.92%	SIM **0.5530** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0469 RSD 18.71%	SIM **0.4092** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.1259 RSD 42.49%	SIM **0.4751** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.1267 RSD 37.70%
RS	SIM **0.4568** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0008 RSD 40.13%	SIM **0.4344** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0058 RSD 43.65%	SIM **0.5139** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0078 RSD 24.46%	SIM **0.2811** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0022 RSD 60.49%	SIM **0.3475** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.0009 RSD 54.43%
SSE	SIM **0.5242** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0666 RSD 31.30%	SIM **0.5021** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0619 RSD 34.87%	SIM **0.3147** $\Delta\mathrm{SIM}_{\mathrm{def}}$ -0.1914 RSD 53.74%	SIM **0.4053** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.1220 RSD 43.04%	SIM **0.4092** $\Delta\mathrm{SIM}_{\mathrm{def}}$ +0.0608 RSD 46.34%

where u is a uniform random perturbation under the same ℓ_∞ bound, and r is a Rademacher sign vector. M3 retains the per-sample absolute perturbation magnitudes before serialization. M4 applies a one-second circular perturbation shift at 24 kHz. Results in Table A3 lead to the following analysis.

Perturbation budget alone does not explain the attack. M1 (Ours) produces the lowest SIM among the ablated versions. M2 under the same ℓ_∞ bound yields a paired gap of $+0.0814$ with 95% CI $[+0.0695, +0.0934]$. Although it also reduces speaker similarity relative to the non-attack control C1, it remains weaker than M1 and does not by itself isolate direction from perturbation energy.

Optimized direction contributes beyond perturbation energy. M3 approximately matches the magnitude and energy of M1: after quantization, the relative ℓ_2 difference is about 6.4×10^{-5}, and the SNR difference is about 0.00055 dB. Despite this energy match, M3 increases SIM from 0.4114 to 0.4535, giving a gap of $+0.0421$ (95% CI $[+0.0319, +0.0525]$). This supports a contribution from the optimized perturbation direction beyond its amplitude and energy.

Temporal alignment also matters. Circularly shifting the optimized perturbation by one second (M4) increases SIM to 0.4789, a gap of $+0.0675$ (95% CI $[+0.0563, +0.0789]$). After PCM_16 processing, the relative Fourier-magnitude difference between M4 and M1 is 9.49×10^{-4}. Hence the weakening of the shifted perturbation supports the importance of its original temporal alignment with the reference speech in this case.

The clean effect of RS. Adding the RS defense reduces clean-condition mean SIM from 0.7750 (Clean) to 0.7426 (C1). The paired difference is -0.0325 (95% CI $[-0.0379, -0.0269]$).

A.5.4 Hyperparameter Sensitivity Analysis

We assess the local sensitivity of JASDA to its two optimization hyperparameters: the number of iterations $\mathcal{K}$ and the step size η in (18). We use the same clean LEMAS-TTS model and the same

Table A3: Ablation analysis for JASDA under the RS defense. The gap value of C1 is contrasted with Clean, while those of M2∼M4 are contrasted with M1 (Ours). A positive value favors M1 (Ours) over M2∼M4. The rightest column counts positive/negative/zero paired gaps. Parentheses contain STDs and 95% CI denotes the 95% confidence interval of Gap w.r.t. the 100 test samples.

Condition	SIM (STD)	Gap	95% CI of Gap	$+/-/0$
Clean	0.7750 (0.0766)	–	–	–
C1	0.7426 (0.0696)	-0.0325	$[-0.0379, -0.0269]$	8/92/0
M1 (Ours)	**0.4114** (0.0729)	0	–	–
M2	0.4929 (0.0832)	$+0.0814$	$[+0.0695, +0.0934]$	90/10/0
M3	0.4535 (0.0857)	$+0.0421$	$[+0.0319, +0.0525]$	79/21/0
M4	0.4789 (0.0779)	$+0.0675$	$[+0.0563, +0.0789]$	90/10/0

100 LibriSpeech-PC test prompts as in Section 4.1 in the no defense case. Following a one-factor-at-a-time protocol, we vary $\mathcal{K} \in \{198, 199, 200, 201, 202\}$ while fixing $\eta = 0.002$, and vary $\eta \in \{0.0018, 0.0019, 0.0020, 0.0021, 0.0022\}$ while fixing $\mathcal{K} = 200$. All other settings are held fixed.

We report SIM and RSD with the fixed clean anchor $\mathrm{SIM}_{\mathrm{clean}} = 0.7750$. We additionally report the paired difference from the regenerated center $\Delta_{\mathrm{ctr}} = \mathrm{SIM}_{\mathrm{setting}} - \mathrm{SIM}_{(200,0.002)}$, together with a paired bootstrap 95% confidence interval (CI) over the 100 test prompts. A negative Δ_{ctr} indicates a lower SIM than the center. The H/L/E column counts samples whose SIM is higher than, lower than, or equal to the center, respectively.

Sensitivity to the number of iterations $\mathcal{K}$ (Table A4). The mean SIM varies from 0.4181 to 0.4279, which gives a narrow range of 0.0098 and indicates the robustness of JASDA to a small change of $\mathcal{K}$. Among the tested values, $\mathcal{K} = 200$ yields the lowest SIM 0.4181. The confidence intervals for $\mathcal{K} \in \{198, 199\}$ include zero, which indicates no clear paired difference from the center at these settings. Since the gap among these settings is small, we set $\mathcal{K} = 200$ as default.

Table A4: Sensitivity of JASDA to $\mathcal{K}$ with η fixed at 0.002.

$\mathcal{K}$	SIM ↓	$\Delta\mathrm{SIM}_{\mathrm{clean}}$	RSD (%)	Δ_{ctr}	95% CI of Δ_{ctr}	H/L/E
198	0.4195	-0.3555	45.88	$+0.001353$	$[-0.003364, +0.006146]$	52/48/0
199	0.4243	-0.3507	45.25	$+0.006215$	$[-0.000521, +0.012977]$	51/49/0
200 (Default)	0.4181	-0.3569	46.05	0	–	–
201	0.4279	-0.3471	44.79	$+0.009814$	$[+0.004396, +0.015414]$	66/34/0
202	0.4234	-0.3516	45.37	$+0.005284$	$[+0.000051, +0.010637]$	57/43/0

Sensitivity to the step size η (Table A5). $\eta = 0.0022$ gives the lowest SIM (0.3698) and the largest RSD (52.29%), while $\eta = 0.0021$ gives the highest SIM (0.4550). The paired confidence intervals exclude zero for $\eta \in \{0.0019, 0.0021, 0.0022\}$. Since all the settings cause significant disruptions, we retain $\eta = 0.002$ as default.

Table A5: Sensitivity of JASDA to η with $\mathcal{K}$ fixed at 200.

η	SIM ↓	$\Delta\mathrm{SIM}_{\mathrm{clean}}$	RSD (%)	Δ_{ctr}	95% CI of Δ_{ctr}	H/L/E
0.0018	0.4238	-0.3512	45.32	$+0.005682$	$[-0.006108, +0.017129]$	49/51/0
0.0019	0.4387	-0.3363	43.39	$+0.020633$	$[+0.008428, +0.032475]$	68/32/0
0.0020 (Default)	0.4181	-0.3569	46.05	0	–	–
0.0021	0.4550	-0.3200	41.29	$+0.036896$	$[+0.025635, +0.048155]$	71/29/0
0.0022	0.3698	-0.4052	52.29	-0.048341	$[-0.060853, -0.035805]$	23/77/0

Overall robustness and interpretation. Across all nine configurations, mean SIM remains between 0.3698 and 0.4550, corresponding to RSD values from 41.29% to 52.29%. Moreover, every

configuration produces a lower SIM than the clean condition for all 100 paired test samples. Thus, the degradation of speaker similarity persists throughout both local neighborhoods.

AI USE STATEMENT

In this work, we used generative AI tools only for grammar correction and minor language edits to improve readability. We have not used generative AI tools for other tasks. We have reviewed all AI-assisted work, and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.