MORE: MULTI-OBJECTIVE ADVERSARIAL ATTACKS ON SPEECH RECOGNITION

Anonymous authors

000

001

002003004

010 011

012

013

014

015

016

017

018

019

020

021

022

024

025

026

027

028

029

031

033

037

038

040

041

042 043

044

046

047

048

051

052

Paper under double-blind review

ABSTRACT

The emergence of large-scale automatic speech recognition (ASR) models such as Whisper has greatly expanded their adoption across diverse real-world applications. Ensuring robustness against even minor input perturbations is therefore critical for maintaining reliable performance in real-time environments. While prior work has mainly examined accuracy degradation under adversarial attacks, robustness with respect to efficiency remains largely unexplored. This narrow focus provides only a partial understanding of ASR model vulnerabilities. To address this gap, we conduct a comprehensive study of ASR robustness under multiple attack scenarios. We introduce MORE, a multi-objective repetitive doubling encouragement attack, which jointly degrades recognition accuracy and inference efficiency through a hierarchical staged repulsion-anchoring mechanism. Specifically, we reformulate multi-objective adversarial optimization into a hierarchical framework that sequentially achieves the dual objectives. To further amplify effectiveness, we propose a novel repetitive encouragement doubling objective (REDO) that induces duplicative text generation by maintaining accuracy degradation and periodically doubling the predicted sequence length. Overall, MORE compels ASR models to produce incorrect transcriptions at a substantially higher computational cost, triggered by a single adversarial input. Experiments show that MORE consistently yields significantly longer transcriptions while maintaining high word error rates compared to existing baselines, underscoring its effectiveness in multi-objective adversarial attack.

1 Introduction

Automatic speech recognition (ASR) models, exemplified by the Whisper family (Radford et al., 2023), have become integral to a wide range of applications, including virtual assistants, real-time subtitling, clinical documentation, and spoken navigation (Gao & Chen, 2024). Despite their success, the reliability of these systems in practical deployments remains fragile: even small adversarial perturbations can substantially degrade recognition accuracy or disrupt inference efficiency—for instance, by causing misinterpretation of user commands or inducing denial-of-service behaviors. These vulnerabilities underscore the need for a systematic examination of ASR robustness across both accuracy and efficiency, which is essential for ensuring dependable performance in real-world, time-sensitive environments.

Most prior work has been dedicated to accuracy robustness under adversarial attacks (Raina et al., 2024; Raina & Gales, 2024; Olivier & Raj, 2022b; Madry et al., 2018a; Dong et al., 2018; Wang & He, 2021; Gao et al., 2024). While these efforts help understanding ASR model accuracy vulnerabilities, the efficiency robustness of ASR models, and their ability to maintain real-time inference under adversarial conditions remain largely unexplored. Such efficiency is critical, as adversaries can exploit it to degrade system responsiveness, e.g., causing systems to output unnaturally long transcripts, severely impacting usability and causing the inference process to be excessively time-consuming. Therefore, enhancing and evaluating the efficiency robustness of ASR models is crucial to ensure their practicality in real-time, user-facing systems.

As efficiency robustness plays a pivotal role in the real-world applicability of deep learning models, there is a growing need to systematically assess it. Recent research has proposed adversarial attack methods to evaluate efficiency robustness in various domains, including computer vision (Li

et al., 2023c; Chen et al., 2022b), machine translation (Chen et al., 2022a), natural language processing (Chen et al., 2023; Li et al., 2023a; Ebrahimi et al., 2018; Li et al., 2019), and speech generation models (Gao et al., 2025). However, research on the efficiency robustness of ASR models under attacks remains critically scarce, with SlothSpeech (Haque et al., 2023) standing as the only known effort. Yet, SlothSpeech does not consider the impact of efficiency attacks on accuracy and does not systematically explore adversarial output patterns. Consequently, this critical dimension of efficiency robustness in ASR models remains insufficiently studied and calls for deeper investigation.

Nevertheless, the robustness of both accuracy and efficiency in ASR models still lags considerably behind human speech recognition performance (Haque et al., 2023; Gao et al., 2024). This stark disparity underscores the need for a more holistic investigation into the vulnerabilities of these models. In this paper, we conduct a comprehensive study of the robustness of the Whisper family, a set of representative large-scale ASR models, with respect to both accuracy and efficiency. To this end, we propose a novel Multi-Objective Repetitive Doubling Encouragement attack approach (MORE) that simultaneously targets both accuracy and efficiency vulnerabilities. Rather than only optimizing towards an accuracy or efficiency objective, the proposed MORE employs an innovative multi-objective (MO) optimization approach to jointly degrade recognition accuracy and inference efficiency through a novel hierarchical repulsion-anchoring optimization strategy within a single network. Motivated by natural human speech repetitions and repetitive decoding loops observed in transformer-based models Xu et al. (2022), we introduce a repetitive encouragement doubling objective (REDO) that promotes duplicative text pattern generation periodically by maintaining accuracy degradation toward producing elongated transcriptions. A novel asymmetric interleaved strategy is introduced to realize the doubling context periodically alongside an end-of-sentence (EOS) suppression objective, which discourages early termination in transcription.

The contributions of this paper include: (a) this paper, for the first time, presents a unified attack approach that jointly targets both accuracy and efficiency robustness against large-scale ASR models in a multi-objective (MO) repulsion-anchoring optimization manner; (b) we propose a novel repetitive encouragement doubling objective (REDO), which bridges efficiency with accuracy gradients via guiding the accuracy-modified gradients towards repetitive elongated semantic contexts, thereby inducing incorrect yet extended transcriptions; and (c) we provide a comprehensive comparative study of diverse attack methods with insightful findings to balance accuracy and efficiency degradation. Extensive experiments demonstrate that the proposed MORE consistently outperforms all baselines in producing longer transcriptions while maintaining strong accuracy attack performance.

2 RELATED WORK

Adversarial Attacks on Speech Recognition. Automatic speech recognition has been extensively studied regarding its vulnerability to attacks. These attacks primarily seek to degrade recognition accuracy by introducing typically subtle perturbations into speech inputs, thereby compromising transcription accuracy (Haque et al., 2023; Olivier & Raj, 2022a; Schönherr et al., 2018; Wang et al., 2022; Zhang et al., 2022; Ge et al., 2023). Notable examples include attacks in the MFCC feature domain (Vaidya et al., 2015; Dan Iter, 2017), targeted attacks designed to trigger specific commands (Carlini et al., 2016), and perturbations constrained to ultrasonic frequency bands (e.g., DolphinAttack Zhang et al. (2017)). Most prior works on attacking ASR have concentrated on traditional architectures, such as CNN or Kaldi-based systems (Wang et al., 2020), with limited exploration into modern large-scale transformer-based ASR models.

Recent advances in ASR have been driven by the emergence of large-scale models, notably OpenAI Whisper (Radford et al., 2023), a transformer-based encoder-decoder architecture trained on large-scale datasets (680K hours of data), demonstrating greater robustness and generalization across diverse speech scenarios. Consequently, there has been an increasing research interest in evaluating the adversarial robustness of Whisper, particularly focusing on accuracy-oriented attacks. Such efforts include universal attacks (Raina et al., 2024; Raina & Gales, 2024), targeted Carlini&Wagner (CW) attacks (Olivier & Raj, 2022b) and gradient-based methods, i.e., projected gradient descent (PGD) (Madry et al., 2018a), momentum iterative fast gradient sign method (MI-FGSM) (Dong et al., 2018), variance-tuned momentum iterative fast gradient sign method (VMI-FGSM) (Wang & He, 2021), as well as speech-aware adversarial attacks (Gao et al., 2024). However, these approaches primarily focus on accuracy robustness, overlooking the critical dimension of model infer-

ence efficiency robustness, which is essential for practical, real-time deployment (Li et al., 2023b). SlothSpeech (Haque et al., 2023) represents the only effort targeting efficiency degradation in ASR models, but it lacks a dual-objective exploration and does not consider structured semantic guidance, limiting its scope and comprehensiveness in assessing adversarial robustness.

Motivations and Applications. Different from previous attacks, our proposed MORE systematically evaluates and undermines both accuracy and efficiency within a single adversarial network, offering a comprehensive understanding of large-scale ASR model's vulnerabilities that previous single-objective methods cannot provide. The significance of studying the adversarial robustness of ASR models, particularly Whisper, is amplified by their potential deployment in hate speech moderation (MacAvaney et al., 2019; Wu & Bhandary, 2020) and private speech data protection. Practically, our proposed MORE can be applied to distort the transcription of harmful or private speech, preventing ASR systems from reliably converting such content into readable text. By inducing incorrect and excessively long transcriptions, MORE effectively obstructs the extraction of sensitive or harmful information, offering a potential mechanism for enhancing speech privacy and safeguarding users from exposure to undesirable content.

3 MORE

3.1 PROBLEM FORMULATION

Victim model. We consider a raw speech input represented as a sequence $X = [x_1, x_2, \ldots, x_T]$. Its corresponding ground-truth transcription is a sequence of text tokens $Y = [y_1, y_2, \ldots, y_L]$. The target ASR model is denoted by a function $f(\cdot)$ that maps a speech sequence to a predicted transcription, i.e., $f(X) = \hat{Y}$. V denotes the model vocabulary (the set of all possible output tokens), and EOS $\in V$ is the end-of-sequence token. Our objective is to construct an adversarial perturbation δ such that the perturbed input $X + \delta$ triggers malicious behavior in the model.

Attack objective. Most existing adversarial attacks on ASR aim solely to maximize transcription error. However, a practically disruptive attack should not only degrade accuracy but also increase latency, thereby impairing the usability of ASR systems in real-time applications. Motivated by this, we formulate the problem as a dual-objective optimization that jointly targets transcription accuracy and computational efficiency:

$$S = \arg \max_{\delta \in \Delta_{\infty}(\epsilon)} (WER(f(X+\delta), Y), |f(X+\delta)|)$$
 (1)

where $\mathrm{WER}(\cdot)$ denotes the word error rate and $|f(\cdot)|$ denotes the length of the predicted sequence. This formulation explicitly seeks perturbations that (1) maximize transcription error relative to the ground truth, and (2) induce excessively long outputs, thereby amplifying computational overhead.

Perturbation constraint. To ensure perturbations remain imperceptible, we impose both energyand peak-based constraints. A standard measure is the signal-to-noise ratio (SNR), which compares the energy of the signal and the perturbation:

$$SNR_{dB} = 20 \log_{10} \left(\frac{\|X\|_2}{\|\delta\|_2} \right). \tag{2}$$

While SNR constrains overall perturbation energy, it may still allow short, high-amplitude distortions. To avoid this, we additionally bound the perturbation's peak amplitude using the ℓ_{∞} norm:

$$\Delta = \{ \delta \mid \|\delta\|_{\infty} \le \epsilon \}, \quad \text{where} \quad \epsilon = \frac{\|X\|_{\infty}}{\text{SNR}}.$$
 (3)

This ℓ_{∞} constraint ensures that no single sample deviates excessively, which aligns with psychoacoustic masking principles. The adversarial example is thus defined as $X_{\text{adv}} = X + \delta$ with $\delta \in \Delta$.

3.2 DESIGN OVERVIEW AND MOTIVATIONS

The proposed **MORE** attack is motivated by the autoregressive nature of ASR models and the different optimization dynamics of our two goals: reducing transcription accuracy and prolonging

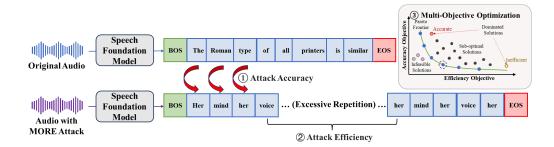


Figure 1: Overview of our proposed multi-objective repetitive encouragement doubling (MORE) adversarial attack method.

decoding for efficiency degradation, as illustrated in Fig. 1. In autoregressive ASR models, each predicted token influences all subsequent predictions, with the end-of-sentence (EOS) token being particularly sensitive—small changes to its logits can dominate the decoding process (Raina et al., 2024; Olivier & Raj, 2022b)—yet the EOS token typically receives only a sparse gradient signal. The accuracy attack objective, in contrast, distributes across many token positions, encouraging mis-transcriptions and resulting in a relatively large feasible adversarial set since many incorrect transcripts are possible. The efficiency attack objective, however, mainly targets the non-stopping behavior associated with a single EOS token, where gradients are narrowly concentrated and typically smaller in magnitude compared with those of the broader accuracy objective. Because of these differences in scale and gradient structure, directly optimizing both objectives together may be difficult to balance in practice. Hence, our proposed MORE attack introduces a hierarchical two-stage repulsion-anchoring attack strategy. The *repulsion* stage establishes a robust mis-transcription by focusing on accuracy degradation, ensuring the model is driven off the correct trajectory. The *anchoring* stage then anchors this solution while applying efficiency attack in the remaining degrees of freedom.

Formally, we approximate the hierarchical formulation in Eq. 4 using a two-stage repulsionanchoring method. In the *repulsion* stage, we maximize a differentiable proxy of WER. In the *anchoring* stage, we extend the decoded sequence length while preserving the high error rate obtained in the *repulsion* stage. The following sections describe the optimization procedure for each in detail.

$$S = \arg \max_{\delta \in \Delta_{\infty}(\epsilon)} \text{WER}(f(X+\delta), Y), \quad \delta^* = \arg \max_{\delta \in S} |f(X+\delta)|, \tag{4}$$

This staged hierarchical design provides a structured way to combine the objectives without forcing them to compete directly.

3.3 REPULSION STAGE: ACCURACY ATTACK

The first pillar of our **MORE** attack is the *repulsion* stage, which focuses on degrading transcription accuracy. Since WER is inherently a non-differentiable metric involving discrete token comparisons, it cannot be optimized directly using gradient-based methods. Therefore, we use a differentiable proxy objective, *i.e.*, the cross-entropy loss to construct the repulsive objective. By minimizing the negative cross-entropy loss between the model's prediction and the ground truth, we effectively guide the model's internal probability distribution away from the correct token sequence. The accuracy attack loss is formulated as:

$$\mathcal{L}_{acc} = -CE(f(X+\delta), Y). \tag{5}$$

Taking gradient changing steps w.r.t. this loss function encourages the ASR model to output incorrect tokens, which directly correlates with an increase in the ultimate WER. This serves as the initial "destabilization" repulsion stage of our attack.

3.4 ANCHORING STAGE: EFFICIENCY ATTACK

Complementing the accuracy degradation, MORE's efficiency attack targets the now-vulnerable model by anchoring it to generate excessively long and computationally expensive transcriptions. This *anchoring* stage is accomplished through the integration of two objectives.

EOS suppression. In most modern speech recognition models, decoding terminates when a special EOS token is predicted. By penalizing the probability of this token, we can deceive the model into prolonging the decoding process indefinitely, often resulting in the generation of irrelevant or meaningless tokens. Penalizing only the EOS token often has limited impact since its probability is typically dominant at the final decoding step. To enhance attack effectiveness, we not only reduce the likelihood of EOS but also explicitly emphasize a competing token, the token with the second largest probability, since reinforcing this alternative both reduces EOS dominance and steers the model toward an alternative continuation. Therefore, the EOS-suppression loss is formulated as:

$$\mathcal{L}_{EOS} = P_L^{EOS} - P_L^z, \tag{6}$$

where $P_L^{\rm EOS}$ is the probability that the model emits EOS at the final position L. z is the token with the second largest probabilities at position L:

$$z = \arg\max_{v \in V \setminus \{\text{EOS}\}} P_L^v \tag{7}$$

We denote L as the output sequence length, and P_L^v is the model's predicted probability of token v being produced at output position L. This dual adjustment ensures the model is discouraged from selecting EOS while being nudged toward an alternative continuation, thereby minimizing this loss reduces EOS dominance and favors continuation tokens, which in turn prolongs decoding.

Repetitive Encouragement Doubling Objective (REDO). While effective, simple EOS suppression can lead to unstable optimization or low-confidence, random outputs. To introduce a more structured and potent method for sequence elongation, we propose a novel *repetitive encouragement doubling objective (REDO)*, inspired by natural speech disfluencies in which human speakers repeat words or phrases. The inspiration for this objective design also comes from sentence-level repetition in transformer-based models (Xu et al., 2022), where sentence-level repetition frequently emerges under maximization-based decoding strategies. In such cases, once a sentence with high generation probability is produced, the model tends to reproduce it in subsequent steps, as its presence in the context further boosts its likelihood of being selected again.

This recursive amplification leads to a self-sustaining loop of repetition, wherein repeated sentences reinforce their own future generation by dominating the context. As identified in (Xu et al., 2022), this self-reinforcing phenomenon causes the model to become trapped in sentence-level repetition loops. Our REDO leverages this mechanism by encouraging the speech recognition model to mimic such structured repetitions, thereby reliably increasing sequence length, as demonstrated in Figure 1. This technique hijacks the model's predictive mechanism, forcing it into a more specific, redundant failure mode. Specifically, given an initial decoding output \hat{Y} , we construct a new target sequence \bar{Y} that contains a repeated segment. We then force the model to predict this new, longer sequence using a cross-entropy objective. The target sequence \bar{Y}_i for step i is constructed as:

$$\bar{Y}_i = \hat{y}_{\lfloor \frac{i}{D} \rfloor} [1:L-1] \parallel \hat{y}_{\lfloor \frac{i}{D} \rfloor} [1:L-1], \tag{8}$$

where L is the length of target sequence length at step $\left\lfloor \frac{i}{D} \right\rfloor$, D is the the doubling period, controlling how frequently the sequence is duplicated. The floor function $\left\lfloor \frac{i}{D} \right\rfloor$ ensures that the repeated segment remains fixed within each interval of D steps, only updating once every D steps. This periodic repetition creates stable semantic loops that encourage longer and more redundant model outputs. The doubling loss REDO is:

$$\mathcal{L}_{REDO} = CE(f(X+\delta), \bar{Y}_i). \tag{9}$$

For a concrete example, if the target sequence for attacking step 0 is $\bar{Y} = [y_1, y_2, y_3, \text{EOS}]$, the target sequence for attacking step 0 to 9 should be $\bar{Y} = [y_1, y_2, y_3, y_1, y_2, y_3]$ with doubling the regular tokens and strictly eliminating the EOS token. This loss explicitly guides the model to produce periodic, repeated segments, which serve to rapidly and reliably inflate the output token count while maintaining a degree of linguistic structure, making the attack more potent. Finally, for efficiency attack, the loss is formulated as $\mathcal{L}_{\text{eff}} = \mathcal{L}_{\text{REDO}} + \mathcal{L}_{\text{EOS}}$.

290

291

292

293

294

295

296

297

298

299

300 301

302

303 304

305 306

307

308

309 310

311

312

313

314

315

316

317

318

319320321

322

323

Algorithm 1 MORE: Hierarchical Attack with Curriculum Interleaved Efficiency Losses

```
271
               1: Input: Original audio X, true transcript Y, \ell_{\infty} radius \epsilon, step size \alpha, total steps K, accuracy steps K_a.
272
                   doubling period D, ASR model f(\cdot).
273
                   Output: Adversarial perturbation \delta
              3: Initialize \delta \leftarrow 0
274
              4: Repulsion Stage: Accuracy (steps 1 \dots K_a)
275
               5: for i = 1 to K_a do
276
                         Compute \mathcal{L} \leftarrow \mathcal{L}_{acc} by Eq. 5
              7:
                         \delta \leftarrow \operatorname{clip}_{[-\epsilon,\epsilon]} (\delta - \alpha \cdot \operatorname{sign}(\nabla_{\delta} \mathcal{L}))
278
               8: end for
              9: Anchoring Stage: Efficiency (steps K_a+1 \dots K)
279
             10: for i = K_a + 1 to K do
280
             11:
                         s \leftarrow i - K_a
281
                         if (s-1) \mod D = 0 then
             12:
282
             13:
                              Calculate \bar{Y} by Eq. 8
283
             14:
                         end if
284
             15:
                         Compute \mathcal{L}_{EOS} by Eq. 6
             16:
                         Compute \mathcal{L}_{RDEO} by Eq. 9
285
             17:
                         \mathcal{L} \leftarrow \mathcal{L}_{EOS} + \mathcal{L}_{REDO}
             18:
                         \delta \leftarrow \text{clip}_{[-\epsilon,\epsilon]} (\delta - \alpha \cdot \text{sign}(\nabla_{\delta} \mathcal{L}))
287
             19: end for
288
             20: return \delta
289
```

Asymmetric interleaving. Applying a single-stage long-repeated target for attack can destabilize gradient optimization with the long-horizon optimization difficulties (Bengio et al., 2009; 2015; Madry et al., 2018b). To mitigate this, REDO breaks the long-horizon repetition task into a sequence of easier subproblems, yielding smoother optimization than trying to force a single-stage long-target objective. REDO is therefore formulated as a stepwise, curriculum-style attack that progressively optimizes for longer repeated outputs. Concretely, we *interleave*: for step s maintain the repeated target fixed when $s \mod D \neq 0$ and extend it to the next longer form when $s \mod D = 0$. This "periodic booster" concentrates high-variance, long-horizon REDO updates sparsely while using frequent short-horizon updates to stabilize learning, and is distinct from summing losses or applying dense REDO updates at every step.

Algorithm Details. We integrate these components into a unified hierarchical procedure (Algorithm 1) to the dual-objective problem. We provide a detailed analysis in Appendix B.

4 EXPERIMENTS

Datasets. We utilize two widely used ASR datasets from HuggingFace, LibriSpeech Panayotov et al. (2015) and LJ-Speech Ito & Johnson (2017), and evaluate the first 500 utterances from each (LJ-Speech and the test-clean subset of LibriSpeech), with all audio resampled to 16,000 Hz.

Experimental Setup. We evaluate five models from the Whisper family (Radford et al., 2023), including Whisper-tiny, Whisper-base, Whisper-small, Whisper-medium, and Whisper-large, all obtained from HuggingFace. In MORE, the hyperparameters I and K_a are set to 10 and 50, respectively. To benchmark the proposed MORE approach, we compare it against five strong baselines: PGD (Olivier & Raj, 2022b), MI-FGSM (Dong et al., 2018), VMI-FGSM (Wang & He, 2021), the speech-aware gradient optimization (SAGO) method (Gao et al., 2024), and SlothSpeech (Haque et al., 2023). To ensure imperceptibility to human listeners, we set the perturbation magnitudes ϵ to 0.002 and 0.0035, which correspond to average signal-to-noise ratios (SNRs) of 35 dB and 30 dB, respectively—both within the range generally considered inaudible to humans (Gao et al., 2024). All experiments are conducted using a single NVIDIA H100 GPU.

Evaluation Metrics. To evaluate ASR accuracy degradation, we adopt word error rate (WER) as the metric, which quantifies the proportion of insertions, substitutions, and deletions relative to the number of ground-truth words. Given that adversarial transcriptions may be excessively long, we truncate the predicted sequence to match the length of the reference text to better evaluate accuracy

	Whisper-tiny		Whisper-base		Whisper-small		Whisper-medium		Whisper-large	
Attack Methods	WER	length	WER	length	WER	length	WER	length	WER	length
LibriSpeech Dataset										
clean	6.66	21.84	4.90	21.77	3.64	21.84	2.99	21.84	3.01	21.84
PGD	93.17	35.02	88.73	31.65	75.23	27.93	64.77	26.94	33.33	21.80
SlothSpeech	46.80	119.38	54.63	156.07	38.25	110.40	31.09	81.75	34.21	79.78
SAGO	93.19	31.93	88.23	33.85	74.67	27.45	62.46	25.97	30.26	21.12
VMI-FGSM	87.91	32.39	93.47	33.90	73.84	27.34	60.74	25.38	29.77	20.79
MI-FGSM	93.32	34.38	87.57	34.44	74.18	27.07	61.16	25.70	34.27	21.40
MORE	91.01	296.28	88.42	300.13	74.28	213.94	64.04	234.25	53.72	301.47
LJ-Speech Dataset										
clean	5.34	18.55	3.77	18.65	3.33	18.64	3.33	18.55	3.36	18.55
PGD	93.63	30.60	90.23	29.98	77.08	22.67	65.07	22.71	26.54	18.55
SlothSpeech	47.10	116.05	59.98	187.9	32.52	65.85	22.23	101.11	21.14	35.33
SAGO	89.53	31.63	84.46	26.69	72.08	22.31	59.08	21.80	21.03	19.11
VMI-FGSM	90.26	29.75	93.85	30.97	75.27	21.90	60.56	21.10	23.00	18.13
MI-FGSM	93.98	32.48	89.14	28.19	74.55	21.80	59.10	22.31	28.02	18.37
MORE	90.85	296.66	89.53	313.64	74.33	208.51	58.86	204.18	43.13	231.52

Table 1: Comparison of average recognition accuracy (WER%) and average transcribed text token length of various attack methods on the LibriSpeech and LJ-Speech datasets at an SNR of 35 dB. The reported accuracy and token length are averaged over 500 utterances for each dataset. 'Clean' denotes performance on the original, unperturbed speech. Note that higher WER and longer transcribed token length indicate a more successful attack.

	Whisper-tiny		Whisper-base		Whisper-small		Whisper-medium		Whisper-large	
Attack Methods	WER	length	WER	length	WER	length	WER	length	WER	length
LibriSpeech Dataset										
clean	6.66	21.84	3.77	21.77	3.64	21.84	2.99	21.84	3.36	21.84
PGD	96.22	35.30	93.90	33.21	85.83	31.40	78.77	28.87	47.93	21.92
SlothSpeech	60.06	123.93	69.33	152.82	56.45	102.15	41.75	77.43	48.34	78.65
SAGO	95.64	34.44	93.72	32.95	83.94	28.28	74.67	27.00	45.67	21.75
VMI-FGSM	93.23	37.53	96.08	37.72	83.15	28.06	73.66	27.02	42.75	22.07
MI-FGSM	96.18	34.37	93.08	33.22	83.49	28.37	72.52	27.54	45.35	22.11
MORE	94.73	300.79	93.70	324.05	86.15	238.52	77.84	202.34	60.90	277.65
LJ-Speech Dataset										
Clean	5.34	18.55	4.90	18.65	3.33	18.64	3.33	18.55	3.01	18.55
PGD	96.56	31.79	94.46	30.50	86.08	25.69	79.45	24.42	42.87	19.97
SlothSpeech	61.62	123.58	71.51	187.87	51.33	76.59	38.33	103.33	35.74	37.50
SAGO	91.23	29.77	87.85	29.67	80.80	22.58	71.50	22.65	33.22	19.08
VMI-FGSM	93.46	28.67	96.57	33.55	84.15	24.15	73.05	22.16	36.28	19.31
MI-FGSM	96.15	31.04	93.37	29.18	83.97	23.37	73.15	22.47	42.71	18.21
MORE	94.80	326.62	94.08	310.57	85.49	222.04	75.01	229.66	54.13	229.08

Table 2: Comparison of average recognition accuracy (WER%) and average transcribed text token length of various attack methods on the LibriSpeech and LJ-Speech datasets at an SNR of 30 dB.

degradation in the initial portion of the output, where meaningful recognition should occur; WER values exceeding 100.00% are capped at 100.00% for normalization. Higher WER indicates lower ASR performance and thus a more effective accuracy attack. Efficiency attack performance is measured by the length of the predicted text tokens, where a greater length indicates a more effective efficiency attack.

5 RESULTS AND DISCUSSION

We evaluate MORE against state-of-the-art baselines on both accuracy and efficiency across multiple ASR models, examine its robustness under different SNR levels, and conduct ablations to analyze component contributions. A case study with adversarial examples and decoded transcriptions further illustrates its effectiveness.

Attack methods	WER	Length	WER	Length
SNR	35	35	30	30
MORE	90.85	296.66	94.80	326.62
MORE - $\mathcal{L}_{ t acc}$	27.58	293.03	34.33	296.91
MORE - $\mathcal{L}_{ t eff}$	93.63	30.60	96.56	31.79
MORE - $\mathcal{L}_{ t EOS}$	93.92	233.84	96.68	269.36
MORE - $\mathcal{L}_{ exttt{REDO}}$	92.42	120.67	95.72	146.54
MORE - $\mathcal{L}_{ t REDO}$ - $\mathcal{L}_{ t acc}$	47.10	116.05	61.62	123.58
MORE - $\mathcal{L}_{ t EOS}$ - $\mathcal{L}_{ t acc}$	6.97	270.21	7.96	307.38
Clean	5.34	18.55	5.34	18.55

Table 3: Ablation study of the proposed MORE approach by removing different components, evaluating adversarial attack performance in terms of accuracy (WER) and efficiency (length) on the LJ-Speech dataset under SNR of 35 dB and 30 dB. The symbol '–' denotes component removal.

5.1 MAIN RESULTS ACROSS DIFFERENT ATTACKS AND MODELS

We compare MORE with SOTA baselines across multiple ASR models, with results on two datasets at 30 dB and 35 dB shown in Table 1 and Table 2.

Our MORE approach consistently achieves superior efficiency attack performance—generating significantly longer transcriptions while maintaining high WER for accuracy degradation across both SNR levels. Specifically, accuracy-oriented baselines (PGD, SAGO, MI-FGSM, and VMI-FGSM) achieve substantially lower transcription lengths (e.g., 31.65 vs. our 300.13), highlighting the effectiveness of our novel repetitive encouragement doubling objective (REDO). Compared to Sloth-Speech, a baseline specifically designed for efficiency attacks, MORE still achieves significantly longer outputs (e.g., 208.52 vs. 65.85), underscoring the efficacy of our doubling loss design in REDO to effectively induce repetitive and extended transcriptions.

Notably, on the robust Whisper-large model, our MORE approach exhibits exceptional performance in both accuracy and efficiency dimensions. It achieves higher accuracy degradation (WER of 53.72 compared to about 30 for accuracy-oriented attack baselines) while producing average transcription lengths of 301.47—roughly 10 times longer than accuracy-oriented baselines and approximately 3.8 times longer than SlothSpeech (79.78). Additionally, SlothSpeech's weaker accuracy degradation (WER of 46.80 vs. MORE's 91.01) highlights the limitations of optimizing solely for efficiency and underscores the necessity of incorporating accuracy objectives for comprehensive attacks. These findings validate the robustness and utility of our proposed multi-objective optimization approach.

5.2 IMPACT OF SNR CONDITIONS

We further investigate the effect of varying SNR levels on attack effectiveness. By comparing results under 30 dB SNR (Table 2) and 35 dB SNR (Table 1), we observe that attacks under 30 dB SNR generally yield stronger performance, characterized by higher WER and significantly longer transcriptions (length). This confirms that lower SNR (i.e., noisier conditions) provides a more favorable environment for adversarial perturbations to succeed.

5.3 ABLATION STUDY

We conduct an ablation study to assess the contribution of each component in the proposed MORE approach (Table 3). Eliminating the accuracy attack loss leads to the collapse of the accuracy attack performance (from 90.85 to 27.58), whereas the efficiency attack remains effective. This indicates that the proposed REDO and EOS objectives are critical for sustaining the efficiency attack. Removing the \mathcal{L}_{EOS} leads to a decline in efficiency performance from 296.66 to 233.84, indicating that the EOS loss facilitates longer output generation by the Whisper model. However, both accuracy (WER > 90) and efficiency (length > 200) attacks remain effective, suggesting that REDO and multi-objective optimization (MO) with accuracy objective are more critical than EOS loss. When REDO is removed, efficiency performance drops drastically from 296.66 to 120.67, highlighting REDO's

Attacks	Adversarial decoded text transcriptions from adversarial speech	Length	WER
Clean	The Roman type of all these printers is similar in character.	12	0.00
PGD	of Rutland who tried before gallopy was ventenies for contemporary priorath probes.	18	100.00
SlothSpeech	The Roman type of all these. The Roman type of all these. We're still in a learning career. We're still in a learning career. We're still in a learning career.	42	72.73
SAGO	the Rolling Tide. For more on our latest ventilators for summary of Fire Esther.	17	90.91
VMI-FGSM	Roman plight, although he's fengomed for the century of my affairs.	14	100.00
MI-FGSM	and Rutland tied for Broadway's ventilence for Senior Order of Fire efforts.	18	100.00
MORE	her mind, her voice, her voice, and	334	100.00

Table 4: Comparison of decoded transcriptions from adversarially generated speech samples across different attacks: clean, PGD, SlothSpeech, SAGO, VMI-FGSM, MI-FGSM and the proposed MORE approaches. Clean represents the original text of the clean speech sample.

central role in promoting structured repetition. Meanwhile, the WER remains above 90, suggesting that sacrificing a small amount of accuracy can substantially benefit efficiency—reflecting the inherent trade-off in balancing the two objectives. Eliminating \mathcal{L}_{acc} in addition to $\mathcal{L}_{\text{REDO}}$ sharply degrades both accuracy attack (WER 90.85 \rightarrow 47.10) and efficiency (length 296.66 \rightarrow 116.05), underscoring the importance of the proposed MO in jointly supporting both attack objectives. Removing both accuracy loss and EOS leads to a WER of 6.97, indicating complete failure of accuracy attacks. However, efficiency remains high (270.21), showing that REDO alone can still sustain efficiency attacks without MO with accuracy loss or EOS. Removing all efficiency attack components—EOS and REDO—causes efficiency to fail completely (length drops to 30.60), with only the accuracy attack (WER 93.63) remaining effective. This confirms that all efficiency design components are essential for achieving successful efficiency degradation. Overall, all components are critical and effectively contribute to the success of the attacks against the victim ASR models.

5.4 CASE STUDY: ADVERSARIAL SAMPLES

To qualitatively demonstrate the effectiveness of MORE, we present decoded adversarial transcriptions in Table 4. Unlike other baselines, which produce either incorrect outputs, MORE generates a fully incorrect transcription with a structured repetition of the sentence "and her voice" over 100 times, resulting in a length of 334 and a WER of 100.00. This showcases the strength of our proposed multi-objective optimization and the repetitive doubling encouragement objective in simultaneously disrupting transcription accuracy and inducing extreme inefficiency through systematic and semantically coherent redundancy. More case studies can be found in Appendix A.

6 CONCLUSION

We propose MORE, a novel adversarial attack approach that introduces multi-objective repulsionanchoring optimization to hierarchically target recognition accuracy and inference efficiency in ASR models. MORE integrates a periodically updated repetitive encouragement doubling objective (REDO) with end-of-sentence suppression to induce structured repetition and generate substantially longer transcriptions while retaining effectiveness in accuracy attacks. Experimental results demonstrate that MORE outperforms existing baselines in efficiency attacks while maintaining comparable performance in accuracy degradation, effectively revealing dual vulnerabilities in ASR models. The code will be made publicly available upon acceptance.

ETHICS STATEMENT

All authors have read and agree to adhere to the ICLR Code of Ethics. We understand that the Code applies to all conference participation, including submission, reviewing, and discussion. This work does not involve human-subject studies, user experiments, or the collection of new personally identifiable information. All evaluations use publicly available research datasets under their respective licenses. No attempts were made to attack deployed systems, bypass access controls, or interact with real users.

Our contribution, **MORE**, is an adversarial method that can degrade both recognition accuracy and inference efficiency of ASR systems. While our goal is to advance robustness research and stress-test modern ASR models, the same techniques maybe misused to (i) impair assistive technologies (e.g., captioning for accessibility), (ii) disrupt safety- or time-critical applications (e.g., clinical dictation, emergency call transcription, navigation), or (iii) increase computational costs for shared services via artificially elongated outputs. We do not provide instructions or artifacts intended to target any specific deployed product or service, and we caution that adversarial perturbations, especially those designed to be inconspicuous, present real risks if applied maliciously.

To reduce misuse risk and support defenders, we suggest that some concrete defenses should be integrated into ASR systems: decoding-time safeguards such as repetition/loop detectors; input-time defenses such as band-limiting; and training-time strategies such as adversarial training focused on EOS/repetition pathologies. We will explore some targeting defense mechanisms against **MORE** in future work.

REPRODUCIBILITY STATEMENT

We take several steps to support independent reproduction of our results. Algorithmic details for **MORE** are provided in Sec. 3 and further clarified in the Appendix B. Dataset choices and preprocessing (LibriSpeech & LJ-Speech, first 500 utterances per set, 16 kHz resampling) are specified in Sec. 4, while exact model variants (Whisper-tiny/base/small/medium/large from HuggingFace), hardware, perturbation budgets/SNRs, and all attack hyperparameters are detailed in Sec. 4. The evaluation protocol is defined in Sec. 4. We will not include the code archive in the submission due to proprietary requirements. Upon acceptance, we will release a public repository mirroring the anonymous package as soon as we get permission.

REFERENCES

- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.
- Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *25th USENIX security symposium* (*USENIX security 16*), pp. 513–530, 2016.
- Simin Chen, Cong Liu, Mirazul Haque, Zihe Song, and Wei Yang. Nmtsloth: understanding and testing efficiency degradation of neural machine translation systems. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1148–1160, 2022a.
- Simin Chen, Zihe Song, Mirazul Haque, Cong Liu, and Wei Yang. Nicgslowdown: Evaluating the efficiency robustness of neural image caption generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15365–15374, 2022b.
- Yiming Chen, Simin Chen, Zexin Li, Wei Yang, Cong Liu, Robby Tan, and Haizhou Li. Dynamic transformers provide a false sense of efficiency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7164–7180, Toronto, Canada, 2023. Association for Computational Linguistics.

- Mike Jermann Dan Iter, Jade Huang. Generating adversarial examples for speech recognition. Stanford Technical Report, 2017.
 - Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9185–9193, 2018.
 - Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018.
 - Xiaoxue Gao and Nancy F Chen. Speech-mamba: Long-context speech recognition with selective state spaces models. In 2024 IEEE Spoken Language Technology Workshop (SLT), pp. 1–8. IEEE, 2024.
 - Xiaoxue Gao, Zexin Li, Yiming Chen, Cong Liu, and Haizhou Li. Transferable adversarial attacks against asr. *IEEE Signal Processing Letters*, 31:2200–2204, 2024.
 - Xiaoxue Gao, Yiming Chen, Xianghu Yue, Yu Tsao, and Nancy F Chen. Ttslow: Slow down text-to-speech with efficiency robustness evaluations. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.
 - Yunjie Ge, Lingchen Zhao, Qian Wang, Yiheng Duan, and Minxin Du. Advddos: Zero-query adversarial attacks against commercial speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 2023.
 - Mirazul Haque, Rutvij Shah, Simin Chen, Berrak Şişman, Cong Liu, and Wei Yang. Slothspeech: Denial-of-service attack against speech recognition models. *Interspeech*, 2023.
 - Keith Ito and Linda Johnson. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.
 - J Li, S Ji, T Du, B Li, and T Wang. Textbugger: Generating adversarial text against real-world applications. In 26th Annual Network and Distributed System Security Symposium, 2019.
 - Yufei Li, Zexin Li, Yingfan Gao, and Cong Liu. White-box multi-objective adversarial attack on dialogue generation. *arXiv preprint arXiv:2305.03655*, 2023a.
 - Yufei Li, Zexin Li, Wei Yang, and Cong Liu. Rt-lm: Uncertainty-aware resource management for real-time inference of language models. *arXiv preprint arXiv:2309.06619*, 2023b.
 - Zexin Li, Bangjie Yin, Taiping Yao, Junfeng Guo, Shouhong Ding, Simin Chen, and Cong Liu. Sibling-attack: Rethinking transferable adversarial attacks against face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24626–24637, 2023c.
 - Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.
 - Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018a.
 - 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*, 2018b.
 - Raphael Olivier and Bhiksha Raj. Recent improvements of asr models in the face of adversarial attacks. *arXiv preprint arXiv:2203.16536*, 2022a.
 - Raphael Olivier and Bhiksha Raj. There is more than one kind of robustness: Fooling whisper with adversarial examples. *arXiv preprint arXiv:2210.17316*, 2022b.

- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *IEEE ICASSP*, pp. 5206–5210, 2015.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.
- Vyas Raina and Mark Gales. Controlling whisper: Universal acoustic adversarial attacks to control multi-task automatic speech recognition models. In 2024 IEEE Spoken Language Technology Workshop (SLT), pp. 208–215. IEEE, 2024.
- Vyas Raina, Rao Ma, Charles McGhee, Kate Knill, and Mark Gales. Muting whisper: A universal acoustic adversarial attack on speech foundation models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7549–7565, 2024.
- Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*, 2018.
- Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine noodles: exploiting the gap between human and machine speech recognition. In 9th USENIX Workshop on Offensive Technologies (WOOT 15), 2015.
- Donghua Wang, Rangding Wang, Li Dong, Diqun Yan, Xueyuan Zhang, and Yongkang Gong. Adversarial examples attack and countermeasure for speech recognition system: A survey. In *International Conference on Security and Privacy in Digital Economy*, pp. 443–468. Springer, 2020.
- Shen Wang, Zhaoyang Zhang, Guopu Zhu, Xinpeng Zhang, Yicong Zhou, and Jiwu Huang. Query-efficient adversarial attack with low perturbation against end-to-end speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 18:351–364, 2022.
- Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1924–1933, 2021.
- Ching Seh Wu and Unnathi Bhandary. Detection of hate speech in videos using machine learning. In 2020 international conference on computational science and computational intelligence (CSCI), pp. 585–590. IEEE, 2020.
- Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. Advances in Neural Information Processing Systems, 35:3082–3095, 2022.
- Chaoning Zhang, Philipp Benz, Adil Karjauv, Jae Won Cho, Kang Zhang, and In So Kweon. Investigating top-k white-box and transferable black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15085–15094, 2022.
- Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 103–117, 2017.

A MORE ADVERSARIAL TRANSCRIPTIONS

To showcase the effectiveness of MORE, we present more decoded adversarial transcriptions in Table 5 and Table 6.

Sample 1	Adversarial decoded text transcriptions from adversarial speech	Length	WER
Clean	in fourteen sixty-five sweynheim and pannartz began printing in the monastery of subiaco near rome	22	0.00
PGD	With a 14-a-tip to the dog, framed by its and clan herbs begin in renting a Ramona scurrying, home-or-dity-actyl-healed blowing.	31	100
SlothSpeech	In 1365. Twain, and patterns. We began printing in the monastery on Superacro. We are going to be here.	32	57.14
SAGO	In recaing 50 5, engraved ideas and claim herbs begin renting a maimanda scurrant you'll be able to keep the app's raw huge role.	32	92.86
VMI-FGSM	The importance in tips to guide, 3D appliance and clan-obes video granting a new monoscaring different types of CDR tools here in the life.	30	100
MI-FGSM	With the important tip to divide, engraved piles and clan herbs begin grinsling a Murmana scurin per Cynthia of Hugh Bully.	29	100
MORE	Say 14-2 to the yard. So, I am glad herbs begin in the printing, they were modest, they're in the home, I just keep the actual yield going. Well, I'm glad. Well	303	100.00
Sample 2	Adversarial decoded text transcriptions from adversarial speech	Length	WER
Clean	they printed very few books in this type, three only; but in their very first books in rome, beginning with the year fourteen sixty-eight	26	0.00
PGD	There's quite a very few dobs on the street. It's very close to New York, so I'm more alanists. If you're in the Euro-Saushen, I'm so creative.	39	100.00
SlothSpeech	They printed very pure robsimbus type. Three only. And they're very first robs in Rome. You're getting with the year 14, 16-8.	31	52.17
SAGO	I print a very few dobs on the scyphs. But it's in there there tortiros you can morelice, within any effort near the coat your rooms might create.	37	52.17
VMI-FGSM	And as front have there, with your bumps and best drives, the case also may fax them their source-of-fit off some well-ince if the payment Okuniro 40 or a safety aid.	37	100.00
MI-FGSM	There's currently as very few dobs on the street flaps in their very tortidoxing Berliners, depending of a year of caffeine I'm so pleased.	33	86.96
MORE	There's quite a very few dubs on the street. It's a great place. You seem very close to York, so I'm more alanists. It's the band-in-the-in-a-day. Are you or I? So, I'm 16, I'm 6, I create. It's a great place. It's	369	100.00

Table 5: Comparison of decoded transcriptions from adversarially generated speech samples across different attacks: clean, PGD, SlothSpeech, SAGO, VMI-FGSM, MI-FGSM and the proposed MORE approaches. Clean represents the original text of the clean speech sample.

Sample 3	Adversarial decoded text transcriptions from adversarial speech	Length	
Clean	and it was a matter of course that in the middle ages, when the craftsmen took care that beautiful form should always be a part of their productions whatever they were	32	
PGD	If Anif wasn't battle with court, he's cut, can him believe dunesque? By McClack to slimped up K-9's ex to use with little pillows to be beldled his beat if lag and bail for gothician in his Trump's Lloyd-Avergen alone.	60	
SlothSpeech	And it was a matter of course, in the middle ages, when the craftsmen took care that beautiful forms should rollers be a part of their productions, whatever they are. But everything. Yeah, you're not(repeat 17 times) Yeah, you're not Yeah, you're not. Yeah, you're not.	114	
SAGO	And until some battle with the Polish scut Hindu-dulisciensky by Mc-Plac slimed some care and ex-eutectful of toarungs, be thrilled he's be a flag of bamper gothish in his tomb, to play out a dang horror.	60	
VMI-FGSM	If anything is in battle with court he's gun kingdom of the Alliance Samuel Pratt has limped up Keanu and read youth with reform on each of the children who has beat up flag on boyfriend on to shamed us and point out your don't you lotta?	55	
MI-FGSM	I don't know if hasn't battled with the poor, Scott, given with the diligence. By McClagg slimps, he hears, ends up doing to a toiling, shrid the soldier, he's beat up fog and bounced her goth or shin his stance and deploy everything, hoda		
MORE	I don't know if I was in battle with court. He's got He's got him in the lead, do you? I'm a clapped, it's slim, it's okay, I'm exactly He's like, it's so own, it's a bit older, he's a bit flack, it's a bit off the gothician, it's strong. What else is this? (repeat 66 times) What else is this? What else is this? What else is this is this is this is this is this is this	363	-
Sample 4	Adversarial decoded text transcriptions from adversarial speech	Length	
Clean	and which developed more completely and satisfactorily on the side of the "lower-case" than the capital letters.	20	
PGD	I've entered Richard of Omen with her work from 2018 and status factor audio. My style Roman roman rumpims Dominic Pappett, all to my idealist.	36	1
SlothSpeech	And which denote more completely and satisfactorily, oh, beside the molecules? You've done the capital(repeat 8 times) You've done the capital	63	
SAGO	and which the Omen of the Fomor from P-plea Mensec satisfactory are postiad Roman root pins. No more powerful of her influence.	31	
VMI-FGSM	If you're a richer dog old, for more can keep these and to have a spare theory, I'm not today Roman Oruk winds. Roman, we're glad to be here.	35	1
MI-FGSM	You've had a rich devolt of her warmth from cleaved deno de téras diasporoia, her stayered Roman rificims. No meekly out the gluder-mervialis.	43]
MORE	I've had a witcher development with her work on 22 and excited as faculty earlier. My style at Roman Roman pimps. Gonna make it happen. I'm all you're glad you're leaving. I'm glad you're leaving (repeat 54 times) I'm glad you're leaving I'm glad you're leaving. I'm glad you're	387	1

Table 6: Comparison of decoded transcriptions from adversarially generated speech samples across different attacks: clean, PGD, SlothSpeech, SAGO, VMI-FGSM, MI-FGSM and the proposed MORE approaches. Clean represents the original text of the clean speech sample.

755

B COMPLEXITY ANALYSIS

Scope and assumptions. We analyze two costs: (i) attack-time compute to craft δ via Algorithm 1, and (ii) victim-time compute when the ASR model decodes on $X+\delta$. Our derivation accounts for encoder self-attention, decoder self-/cross-attention, feed-forward layers, vocabulary projection/softmax, greedy decoding used to materialize doubled targets, and the backward/forward constant factor. We express totals in terms of model hyperparameters and the scheduling parameters (K, K_a, D) defined in Algorithm 1, and we refer to the objectives in Eqs. 5, 6, 8, and 9.

Notation. Let F be the number of encoder time steps (e.g., log-Mel frames). Encoder depth/width/FF width/heads are $N_e, d_e, d_{\mathrm{ff},e}, h_e$; decoder counterparts are $N_d, d_d, d_{\mathrm{ff},d}, h_d$. Vocabulary size is V. The total PGD steps are K, Stage 1 steps are K_a , and the Stage 2 block length is D (Algorithm 1). Stage 2 has $M = \lceil (K - K_a)/D \rceil$ blocks indexed by $m = 1, \ldots, M$. We denote by B_m the base segment (non-EOS) used to build the doubled target in block m, and by $L_m = |B_m|$ its length. The doubly repeated target in block m has length $\bar{\ell}_m = 2L_m$. We write $\kappa \in [2,3]$ (empirically range) for the backward/forward multiplier. Let T be the number of raw samples in X (used only for the O(T) PGD update).

On the construction of doubled targets. Eq. 8 defines \bar{Y}_i via $\hat{y}_{\lfloor \frac{i}{D} \rfloor}$ but does not specify how the hypothesis \hat{Y} is obtained at the start of each block. To make Eq. 8 operational, we explicitly realize \hat{Y} with a *greedy decode* on the current perturbed input:

$$\hat{Y}^{(m)} = \text{GreedyDecode}(f, X + \delta), \tag{10}$$

$$B_m = \operatorname{STRIPEOS}(\hat{Y}^{(m)}), \tag{11}$$

$$\bar{Y}^{(m)} = B_m \parallel B_m, \tag{12}$$

$$\bar{\ell}_m = 2|B_m|. \tag{13}$$

Here, Eq. 10 computes decoding the perturbed input $X+\delta$ with the victim model $f(\cdot)$ using greedy decoding (selecting the most probable token at each step until termination). In Eq. 11, STRIPEOS(\cdot) removes the terminal EOS token from the decoded hypothesis, leaving only the content-bearing tokens. The notation " $\|$ " denotes sequence concatenation, so Eq. 12 constructs the doubled sequence by repeating B_m back-to-back. Finally, $\bar{\ell}$ calculated by Eq. 13 is the length of this doubled target. This makes Eq. 8 explicit and adds a per-block greedy-decoding cost accounted for below.

Per-pass building blocks. We use standard Transformer accounting; QKV and output projections are absorbed in big- \mathcal{O} terms. For a decoder sequence length ℓ ,

$$C_{\rm enc}(F) = \mathcal{O}\left(N_e\left(F^2d_e + Fd_ed_{\rm ff,e}\right)\right),\tag{14}$$

$$C_{\text{dec-TF}}(\ell, F) = \mathcal{O}\left(N_d(\ell^2 d_d + \ell F d_d + \ell d_d d_{\text{ff},d})\right),\tag{15}$$

$$C_{\text{dec-gen}}(\ell, F) = \mathcal{O}\left(N_d(\ell^2 d_d + \ell F d_d + \ell d_d d_{\text{ff},d})\right). \tag{16}$$

The encoder cost in Eq. 14, $C_{\rm enc}(F)$, measures the cost of processing F input frames, scaling as F^2d_e for self-attention and $Fd_ed_{\rm ff,e}$ for feed-forward layers over N_e encoder layers. The decoder cost under teacher forcing in Eq. 15, $C_{\rm dec\text{-}TF}(\ell,F)$, captures the cost of processing a target sequence of length ℓ , scaling quadratically in ℓ from decoder self-attention, linearly in ℓF from cross-attention with encoder outputs, and linearly in $\ell d_{\rm ff,d}$ from feed-forward layers over N_d decoder layers. The decoder cost under generation in Eq. 16, $C_{\rm dec\text{-}gen}(\ell,F)$, has the same asymptotic form as teacher forcing since autoregressive decoding still requires self-attention, cross-attention, and feed-forward passes, though key-value caching can reduce constants in practice. In both modes, an additional $\mathcal{O}(\ell V)$ term arises from vocabulary projection and softmax, which is significant for large vocabularies but dominated by ℓ^2 self-attention when ℓ is large. Both Eq. 15 and Eq. 16 scale as ℓ^2 (self-attention) and as ℓF (cross-attention). The vocabulary projection/softmax adds $\mathcal{O}(\ell V)$ per forward/backward; we keep it explicit when informative.

Loss conditioning and reuse of passes. Eqs. 5 and 9 are cross-entropy (CE) objectives and must be computed under *teacher forcing* to provide stable gradients. We assume CE terms use teacher forcing throughout. For Eq. 6, we define P_L^v at the *last teacher-forced position* (so $L = \bar{\ell}_m$ in Stage 2). With Algorithm 1 summing $\mathcal{L}_{EOS} + \mathcal{L}_{REDO}$ each Stage 2 step, both losses share *one* forward/backward pass at length $\bar{\ell}_m$; no additional pass is required for EOS.

ATTACK-TIME COMPLEXITY

 Stage 1 (Accuracy; K_a steps). Each step backpropagates \mathcal{L}_{acc} (Eq. 5) under teacher forcing on Y of length $L_{acc} = |Y|$:

$$C_{\text{step}}^{(1)} = \kappa [C_{\text{enc}}(F) + C_{\text{dec-TF}}(L_{\text{acc}}, F)] + \mathcal{O}(L_{\text{acc}}V). \tag{17}$$

Optional early-stopping evaluations (e.g., greedy WER every E steps) add

$$C_{\text{eval,S1}} \approx \left[\frac{K_a}{E}\right] \cdot \left(C_{\text{enc}}(F) + C_{\text{dec-gen}}(L_{\text{eval}}, F) + \mathcal{O}(L_{\text{eval}}V)\right),$$
 (18)

where L_{eval} is the decoded length at evaluation.

Stage 2 (Efficiency; $K-K_a$ steps). Stage 2 consists of M blocks of D steps. In each block m:

• Greedy anchor (once per block). Build $\bar{Y}^{(m)}$ by decoding $\hat{Y}^{(m)}$ and doubling (Eq. 8):

$$C_{\text{greedy},m}^{(2)} = C_{\text{enc}}(F) + C_{\text{dec-gen}}(L_m^{\text{gen}}, F) + \mathcal{O}(L_m^{\text{gen}}V), \quad \text{with } L_m^{\text{gen}} \approx L_m. \tag{19}$$

• PGD steps (every step in the block). Algorithm 1 uses the sum $\mathcal{L}_{EOS} + \mathcal{L}_{REDO}$ each step, with teacher-forced length $\bar{\ell}_m = 2L_m$. Hence, per step:

$$C_{\text{step}}^{(2)}(m) = \kappa \left[C_{\text{enc}}(F) + C_{\text{dec-TF}}(\bar{\ell}_m, F) \right] + \mathcal{O}(\bar{\ell}_m V), \tag{20}$$

and over D steps:

$$C_{\text{block}}^{(2)}(m) = D C_{\text{step}}^{(2)}(m) = D \kappa [C_{\text{enc}}(F) + C_{\text{dec-TF}}(2L_m, F)] + \mathcal{O}(D 2L_m V).$$
 (21)

Summing over blocks,

$$C_{\text{Stage2}} = \sum_{m=1}^{M} \left(C_{\text{greedy},m}^{(2)} + C_{\text{block}}^{(2)}(m) \right)$$

$$= \underbrace{\kappa(K - K_a) C_{\text{enc}}(F)}_{\text{encoder re-run each PGD step}} + \kappa N_d \sum_{m=1}^{M} \left[D \left(2L_m \right)^2 d_d + D \left(2L_m \right) \left(F d_d + d_d d_{\text{ff},d} \right) \right]$$

$$+ \sum_{m=1}^{M} C_{\text{greedy},m}^{(2)} + \mathcal{O}\left(V \sum_{m=1}^{M} D \left(2L_m + L_m^{\text{gen}} \right) \right). \tag{22}$$

Growth envelopes for L_m . The doubled-target curriculum encourages L_m to increase across blocks.

• Geometric (until cap). In this case, L_m grows by doubling until it reaches the cap L_{max} :

$$L_m = \min\{L_0 \, 2^{m-1}, \, L_{\text{max}}\},\tag{23}$$

$$M^* = \min \left\{ M, 1 + \lfloor \log_2(L_{\text{max}}/L_0) \rfloor \right\}. \tag{24}$$

Then the sums are

$$\sum_{m=1}^{M} L_m = L_0(2^{M^*} - 1) + (M - M^*) L_{\text{max}},$$
 (25)

$$\sum_{m=1}^{M} L_m^2 = \frac{L_0^2}{3} (4^{M^*} - 1) + (M - M^*) L_{\text{max}}^2.$$
 (26)

Plugging into Eq. 22, the self-attention term scales as $\Theta(D \cdot 4^{M^*})$ before saturation at L_{max} .

• Linear (until cap). If $L_m = L_0 + (m-1)\Delta$ (capped at L_{max}), then the uncapped sums are

$$\sum_{m=1}^{M} L_m = \frac{M}{2} (2L_0 + (M-1)\Delta), \tag{27}$$

$$\sum_{m=1}^{M} L_m^2 = ML_0^2 + L_0 \Delta M(M-1) + \frac{\Delta^2}{6} (M-1)M(2M-1).$$
 (28)

Here the dominant term in the self-attention sum is $\Theta(D\,M^3\Delta^2)$ when $\Delta>0$.

Total attack-time cost. Combining stages,

$$C_{\text{attack}} = \kappa K_a [C_{\text{enc}}(F) + C_{\text{dec-TF}}(L_{\text{acc}}, F)] + C_{\text{eval,S1}} + C_{\text{Stage2}}, \tag{29}$$

with C_{Stage2} calculated from Eq. 22. The PGD update is $\mathcal{O}(T)$ and negligible.

VICTIM-TIME (INFERENCE) COMPLEXITY

When decoding on $X+\delta$ without backprop, expected per-example compute is

$$C_{\text{infer}}(\ell_{\text{adv}}) = C_{\text{enc}}(F) + C_{\text{dec-gen}}(\ell_{\text{adv}}, F) + \mathcal{O}(\ell_{\text{adv}}V),$$
 (30)

where ℓ_{adv} is the induced output length under the chosen decoding policy (greedy by default). Focusing on decoder-dominant terms, the slowdown relative to the clean case (ℓ_{clean}) is

$$\frac{C_{\rm infer}(\ell_{\rm adv})}{C_{\rm infer}(\ell_{\rm clean})} \, \approx \, \frac{\ell_{\rm adv}^2 + \ell_{\rm adv} F + \ell_{\rm adv} \frac{d_{\rm ff,\it d}}{d_{\it d}}}{\ell_{\rm clean}^2 + \ell_{\rm clean} F + \ell_{\rm clean} \frac{d_{\rm ff,\it d}}{d_{\it d}}}, \label{eq:clean_control}$$

highlighting quadratic sensitivity to $\ell_{\rm adv}$ from decoder self-attention. Under the geometric envelope above, $\ell_{\rm adv}$ can grow proportionally to $2^{M^\star}L_0$ until capped by the implementation's maximum length, potentially yielding a $\Theta(4^{M^\star})$ increase in decoder FLOPs before saturation.

Summary bound. With $M = \lceil (K - K_a)/D \rceil$ and $\bar{\ell}_m = 2L_m$, the attack-time compute admits

$$C_{\text{attack}} \in \mathcal{O}\left(\kappa K C_{\text{enc}}(F) + \kappa N_d D \sum_{m=1}^{M} \left(\underbrace{\bar{\ell}_m^2}_{4\bar{L}_m^2} d_d + \bar{\ell}_m \left(F d_d + d_d d_{\text{ff},d}\right)\right) + \sum_{m=1}^{M} C_{\text{greedy},m}^{(2)} + V \cdot D \sum_{m=1}^{M} \bar{\ell}_m\right).$$

$$(31)$$

This bound separates: (i) repeated encoder passes linear in K and F; (ii) decoder self-attention terms growing with $\sum_{m} L_{m}^{2}$ (the principal driver under elongation); and (iii) vocabulary and greedy-decoding overheads linear in sequence length and V.

C THE USE OF LARGE LANGUAGE MODELS

We used a large language model (ChatGPT) solely as a writing assist tool for grammar checking, wording consistency, and style polishing of author-written text. All technical content, results, and conclusions originate from the authors. Suggested edits were reviewed by the authors for accuracy and appropriateness before inclusion. No confidential or proprietary data beyond manuscript text was provided to the tool. This disclosure is made in accordance with the venue's policy on LLM usage.