Boosting ASR Robustness via Test-Time Reinforcement Learning with Audio-Text Semantic Rewards

Linghan Fang[1,2], **Tianxin Xie**[1], **Li Liu**[1]*

[1]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2]Technical University of Munich, Munich, Germany
linghan.fang@tum.de, txie151@connect.hkust-gz.edu.cn, avrillliu@hkust-gz.edu.cn

Abstract

Recently, Automatic Speech Recognition (ASR) systems (e.g., Whisper) have achieved remarkable accuracy improvements but remain highly sensitive to real-world unseen data (data with large distribution shifts), including noisy environments and diverse accents. To address this issue, test-time adaptation (TTA) has shown great potential in improving the model adaptability at inference time without ground-truth labels, and existing TTA methods often rely on pseudo-labeling or entropy minimization. However, by treating model confidence as a learning signal, these methods may reinforce high-confidence errors, leading to confirmation bias that undermines adaptation. To overcome these limitations, we present **ASR-TRA**, a novel **T**est-time **R**einforcement **A**daptation framework inspired by causal intervention. More precisely, our method introduces a learnable decoder prompt and utilizes temperature-controlled stochastic decoding to generate diverse transcription candidates. These are scored by a reward model that measures audio-text semantic alignment, and the resulting feedback is used to update both model and prompt parameters via reinforcement learning. Comprehensive experiments on LibriSpeech with synthetic noise and L2 Arctic accented English datasets demonstrate that our method achieves higher accuracy while maintaining lower latency than existing TTA baselines. Ablation studies further confirm the effectiveness of combining audio and language-based rewards, highlighting our method's enhanced stability and interpretability. Overall, our approach provides a practical and robust solution for deploying ASR systems in challenging real-world conditions.

Code — https://github.com/fangcq/ASR-TRA

Introduction

Recent advances in automatic speech recognition (ASR) have been driven by breakthroughs in self-supervised learning and large-scale weakly supervised training (Baevski et al. 2020; Hsu et al. 2021; Schneider et al. 2019), which enables models to learn rich acoustic and linguistic representations from vast amounts of unlabeled or loosely labeled speech data. Leveraging these techniques, models such as wav2vec 2.0 (Baevski et al. 2020) and Whisper (Radford et al. 2023) have significantly improved transcription accuracy and generalization in various domains.

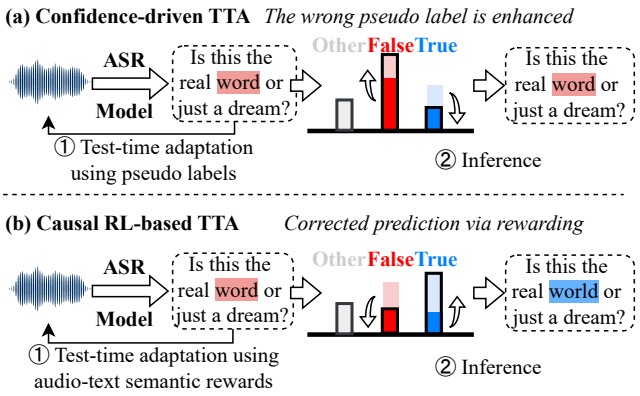

Figure 1: **Overview of test-time adaptation strategies under noisy conditions.** (a) Confidence-driven adaptation reinforces incorrect but high-probability predictions (e.g., *word*), leading to persistent errors. (b) Our reward-guided adaptation **ASR-TRA** favors semantically accurate alternatives (e.g., *world*) through reinforcement learning, even when their initial confidence is low.

Despite these improvements, deploying ASR in real-world applications such as edge devices, online streaming, and resource-constrained small-model scenarios remains challenging. In practice, lightweight ASR systems often face severe out-of-distribution (OOD) conditions (Hendrycks and Dietterich 2019), including background noise, heavy accents, and regional dialects, which are under-represented in training data and lead to significant domain shifts (Ben-David et al. 2007). Traditional methods to improve robustness typically rely on offline retraining or supervised domain adaptation, such as multi-condition training or noise augmentation (Ko et al. 2015; Li, Deng, and Gong 2020), but these approaches are infeasible during test time because labeled data is unavailable.

In the literature, test-time adaptation (TTA) has emerged as a promising approach to enhance ASR robustness without requiring additional training or labeled supervision (Wang

*Corresponding author.

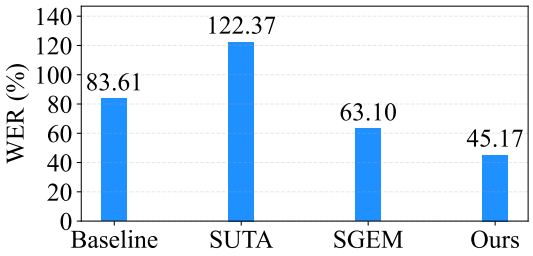

Figure 2: We evaluate WER on the top-100 high-confidence samples from LibriSpeech test-other corrupted with Gaussian noise. While baseline performance from Whisper-Tiny suffers under noise, heuristic methods like SUTA further degrade due to overconfidence. In contrast, ours achieves lower WER. (Lower is better.)

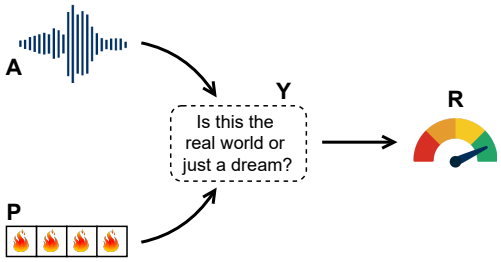

Figure 3: Structural Causal Model (SCM) schematic. Nodes A, P, Y, and R denote audio features, learnable prompt, transcription output, and reward, respectively, with causal flow $A, P \rightarrow Y \rightarrow R$ (Pearl 2009).

et al. 2021; Lin, Li, and Lee 2022; Kim et al. 2023). By enabling models to adapt dynamically to test-time inputs, TTA helps mitigate the mismatch between training and deployment conditions. However, current TTA methods have the following two main challenges, when directly apply to ASR systems (see Fig. 1). **Challenge 1:** Although Whisper is a powerful encoder-decoder model, it lacks a Whisper-specific adaptation mechanism (Radford et al. 2023). **Challenge 2:** Many existing approaches rely on heuristic optimization strategies and heavily use self-generated pseudo-labels (Wang et al. 2021; Lin, Li, and Lee 2022; Lee 2013; Sohn et al. 2020), which can be slow, unstable in acoustically diverse environments, and prone to compounding errors when the model is uncertain. As illustrated in Figure 2, confidence-based methods such as SUTA (Lin, Li, and Lee 2022) can perform even worse on high-confidence samples under noise, highlighting the misalignment between confidence and true accuracy that motivates our approach (see the Experiments section for details).

To address these challenges, we present a novel method **ASR-TRA** (**ASR** with **T**est-time **R**einforcement **A**daptation), which adopts a reinforcement learning (RL) (Williams 1992; Ziegler et al. 2019) perspective and frames TTA as a reward-driven decision process under uncertainty. ASR-TRA addresses these limitations based on two key ideas. **Idea 1:** To leverage the structure of encoder-decoder architectures such as Whisper, we introduce a dedicated adaptation pathway through a learnable decoder prompt, enabling efficient and low-overhead test-time optimization tailored to ASR. **Idea 2:** Instead of relying on fragile heuristics or self-confirming pseudo-labels, we introduce an external semantic reward to guide adaptation more robustly. This avoids compounding errors and allows for more stable updates without ground-truth supervision.

To realize these two key ideas, ASR-TRA begins with a structured modeling of the adaptation process using a Structural Causal Model (SCM) (Pearl 2009), which bridges the motivation with concrete implementation. The SCM comprises four key variables: the encoded audio features A, the learnable decoder prompt P, the generated transcription Y, and the reward R. The directed edges capture the causal dependencies among these components (see Figure 3).

This model not only clarifies how adaptation unfolds in our framework, but also guides the design of three concrete components that operationalize reward-driven learning and prompt-based adaptation at test time: 1) **Prompt injection (for Idea1):** Insert a learnable vector P at the beginning of the decoder input sequence, so that P is processed alongside the token embeddings and directly influences the decoding process as a causal intervention. 2) **Candidate generation & evaluation (for Idea2):** Adjust the sampling temperature to produce multiple diverse transcription hypotheses. Each candidate is then evaluated using CLAP (Contrastive Language–Audio Pretraining) (Elizalde et al. 2023), which can compute an audio–text similarity score as a sequence-level reward to guide adaptation. 3) **Parameter update:** Apply a policy-gradient RL algorithm to backpropagate CLAP rewards and jointly update both the prompt parameters P and the model weights, steering future outputs toward higher reward.

In summary, the proposed ASR-TRA approach offers three main contributions:

- We formulate TTA as an RL process guided by the audio-text reward model CLAP, which mitigates error accumulation from heuristic pseudo-label or confidence-based methods.
- We design a Whisper-specific Structural Causal Model (SCM) where a learnable decoder prompt modulates the decoding process. Combined with policy-gradient updates and the CLAP reward, this framework enables a principled and lightweight TTA approach for ASR.
- Experiments with noisy and accented speech benchmarks show that it consistently outperforms previous TTA methods in both speed and recognition accuracy.

Related Work

Automatic Speech Recognition. Recent progress in automatic speech recognition has been largely driven by the development of large-scale neural network models trained on diverse audio-text corpora. Wav2vec 2.0 (Baevski et al. 2020) leverages self-supervised contrastive pretraining to learn robust speech representations from unlabeled audio. Whisper (Radford et al. 2023), built on a transformer

encoder-decoder architecture, achieves strong multilingual transcription by training on 680k hours of weakly labeled data. Despite their impressive performance, these models remain sensitive to distribution shifts, such as environmental noise, speaker accents, or spontaneous speech, underscoring the need for adaptive mechanisms (Hendrycks and Dietterich 2019; Ben-David et al. 2007) that can generalize well to real-world conditions.

Causality in Deep Learning. Causal reasoning offers a principled framework for improving model robustness and interpretability. Classical formulations (Pearl 2009) define interventions as mechanisms for actively altering model variables to study downstream effects. Recent work has explored causal representation learning (Schölkopf et al. 2021; Mitrovic et al. 2021), which aims to disentangle generative factors of variation for greater robustness under domain shift. In deep learning, causal interventions can be realized through architectural modifications or controlled sampling schemes and have been shown to help models adapt more reliably to distributional changes.

Test-Time Adaptation in ASR. TTA enables models to adjust during inference using only unlabeled test data. Among early approaches, Tent (Wang et al. 2021) proposed an *episodic* TTA framework that adapts the model to each test sample by minimizing entropy while freezing most parameters. This inspired a range of episodic TTA methods that aim for rapid and lightweight domain adaptation. In the ASR domain, SUTA (Lin, Li, and Lee 2022) adapts Whisper-style models by generating pseudo-labels from high-confidence predictions, while SGEM (Kim et al. 2023) minimizes sequence-level entropy to regularize the output distribution. These methods, however, rely heavily on model-internal signals such as entropy or confidence, which can become unreliable under severe distribution shift. Furthermore, they lack a mechanism to incorporate feedback external to the model's prior beliefs.

Reinforcement Learning for TTA. RL offers a general framework for optimizing decisions based on delayed feedback (Sutton and Barto 2018), and has been widely applied to robotics (Levine et al. 2016), dialogue (Li et al. 2016), and language model alignment (Ouyang et al. 2022; Ziegler et al. 2019). Recently, RL has also been explored in test-time adaptation. For instance, RLCF (Zhao et al. 2024) learns reward-driven policies for classifier adaptation. More recent efforts, such as BiTTA (Lee et al. 2025), leverage binary feedback (correct/incorrect) to guide episodic adaptation. While promising, these methods have been mostly applied to classification or navigation tasks, where the action space and feedback signals are comparatively straightforward. In contrast, their application to structured sequence modeling problems such as ASR remains limited.

Positioning of ASR-TRA. While prior TTA methods in ASR, such as SUTA and SGEM, rely on internal heuristics like confidence and entropy, they often lack robustness under distribution shift and may suffer from pseudo-label feedback loops. Recent RL-based approaches such as BiTTA demonstrate the utility of reward signals, but are designed for classification or navigation tasks and have not been adapted for structured sequence modeling in ASR. ASR-TRA bridges this gap by combining causal interventions with external reward-guided adaptation in a Whisper-based framework. Specifically, we introduce learnable decoder prompts as causal variables and optimize them using semantic feedback (Elizalde et al. 2023), enabling robust, interpretable, and efficient adaptation without relying on internal certainty estimates.

Method

We propose a TTA framework, ASR-TRA, for Whisper that integrates causal reasoning with reinforcement learning. The key insight is to treat prompt injection in the Whisper decoder as a causal intervention and the decoding process as a generator of counterfactual hypotheses. We optimize the model at inference time by rewarding generations that better align with CLAP's predictions, without relying on ground-truth transcripts. This enables dynamic and label-free adaptation to unseen acoustic conditions.

Whisper Architecture

The input speech signal is first preprocessed and converted into a log Mel-spectrogram $s \in \mathbb{R}^{F \times T}$ with F frequency bins and T time frames (Logan 2000; McFee et al. 2015). Whisper is a Transformer-based encoder–decoder model (Vaswani et al. 2017; Radford et al. 2023) that follows an autoregressive sequence-to-sequence formulation widely used in ASR (Chan et al. 2016). The audio encoder $\mathrm{Enc}(\cdot)$ produces a hidden representation:

$$h = \mathrm{Enc}(s), \tag{1}$$

and the text decoder generates the output sequence $\hat{y} = (y_1, y_2, \ldots, y_N)$ by modeling the conditional token distribution over output probabilities:

$$P(y_t \mid y_{<t}, h) = \mathrm{Dec}(y_{<t}, h), \tag{2}$$

where $y_{<t}$ denotes previously emitted tokens. In practice, the decoding process can be performed using greedy selection, beam search, or stochastic sampling regulated by temperature scaling (Holtzman et al. 2020).

Prompt Injection as Causal Intervention

We introduce a learnable prompt vector $p \in \mathbb{R}^{L \times d}$, where L denotes the number of prompt tokens and d is the decoder embedding dimension. A soft prompt is prepended to the decoder's input embeddings, directly concatenated before the embedding of the <bos> token. During generation, the decoder attends to both this prompt and the encoder output, allowing the prompt to guide each prediction instead of relying solely on autoregressive decoding from scratch.

Because the prompt is visible to the decoder's attention at every step, it can directly shape the hidden states and thus influence all subsequent token predictions (Li and Liang 2021; Lester, Al-Rfou, and Constant 2021).

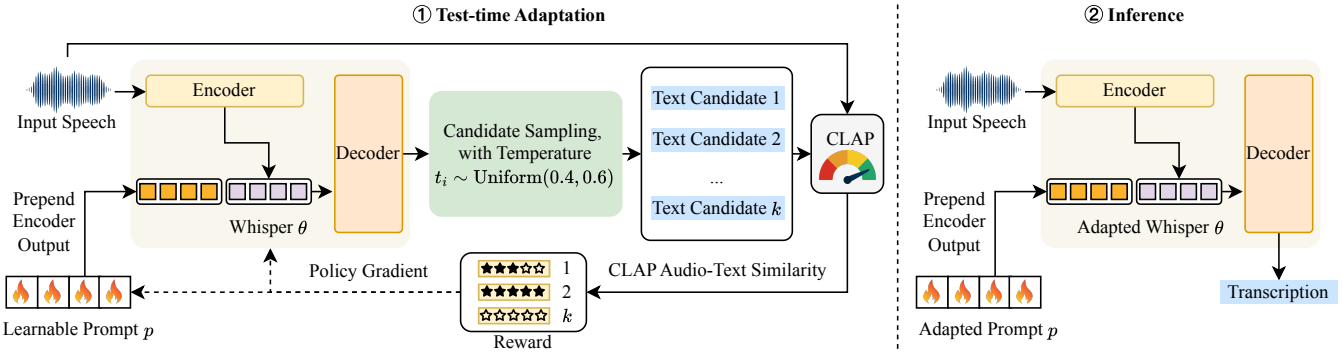

Figure 4: Test-time self-adaptation for Whisper. A baseline transcript is decoded from the input Mel-spectrogram; soft-prompted variants are then sampled at varied temperatures, CLAP scores them all, and the aggregated rewards update the model, achieving on-the-fly correction without labels.

Crucially, we treat p not merely as an additional condition but as a *causal intervention* on the generation process. Formally, the decoder prediction can be written as

$$y_t = \text{Dec}(h, y_{<t}, \text{do}(p)), \tag{3}$$

where $\text{do}(p)$ follows Pearl's do-calculus (Pearl 2009) and indicates that p is set externally rather than inferred from the observed input s. This intervention perturbs the internal generation dynamics without modifying the acoustic input, enabling the model to explore alternative hypotheses under the same observation.

Counterfactual Sampling and Reward Evaluation

To explore diverse output trajectories under a fixed input and prompt, we adopt stochastic decoding by sampling tokens with a temperature parameter $T > 0$ (Holtzman et al. 2020):

$$P_T(y_t \mid y_{<t}, h, p) \propto \exp\left(\frac{\log P(y_t \mid y_{<t}, h, p)}{T}\right). \tag{4}$$

A higher temperature T flattens the token distribution and encourages diversity, while $T \to 0$ approaches greedy decoding. By sampling repeatedly, we obtain K candidate transcriptions $\{\hat{y}^{(1)}, \ldots, \hat{y}^{(K)}\}$, each representing a *counterfactual hypothesis*, i.e., a plausible alternative transcription trajectory conditioned on the same audio and prompt.

Each sampled hypothesis is then evaluated using a reward function $R(\hat{y}^{(i)}) \in \mathbb{R}$ that quantifies its semantic alignment with the input. In our implementation, we adopt the CLAP audio–text similarity model (Elizalde et al. 2023) as the primary reward, which computes the cosine similarity between audio embeddings of input speech and text embeddings of generated transcriptions:

$$r^{(i)} = R(\hat{y}^{(i)}). \tag{5}$$

We also use the scores from other pretrained language models (LM) (Ziegler et al. 2019; Ouyang et al. 2022) as a supplementary signal in our ablation study (Table 3).

This scalar feedback serves as a proxy for transcription quality and enables label-free optimization. The resulting rewards are aggregated and used to update both the prompt vector p and model parameters via a policy-gradient objective (Williams 1992; Sutton and Barto 2018), thereby biasing the model towards generations that achieve higher reward under the same conditions.

Optimization via Reinforcement Learning

To adapt the prompt vector p and model parameters online, we formulate a reinforcement learning objective that explicitly encourages generations with higher semantic quality as measured by the reward function. We treat the prompt-conditioned Whisper as a stochastic policy $\pi_p(\hat{y})$ over output sequences $\hat{y}$ and aim to maximize the expected reward:

$$\mathcal{J}(p) = \mathbb{E}_{\hat{y} \sim \pi_p}[R(\hat{y})]. \tag{6}$$

This setting naturally aligns with policy-gradient methods (Sutton and Barto 2018). We apply the classic REINFORCE algorithm (Williams 1992) to estimate gradients of $\mathcal{J}(p)$. For a batch of N sampled hypotheses $\{\hat{y}^{(i)}\}_{i=1}^{N}$ with corresponding rewards $r^{(i)}$ and log-probabilities $\log P(\hat{y}^{(i)})$, we introduce a baseline to reduce gradient variance. The baseline is simply the mean reward across the batch:

$$\bar{r} = \frac{1}{N} \sum_{i=1}^{N} r^{(i)}. \tag{7}$$

Using this baseline, the gradient of the RL objective with respect to the learnable prompt p and Whisper's parameters θ can be estimated as follows:

$$\nabla_{\theta,p} \mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta,p} \log P(\hat{y}^{(i)}) \cdot (r^{(i)} - \bar{r}), \tag{8}$$

which increases the likelihood of high-reward generations while penalizing low-reward alternatives.

During test time, we perform optimization independently for each input sample, using a small number of reward-evaluated candidates to update the prompt and model parameters. Once the adaptation and prediction for the current sample is completed, the model parameters are restored to their original state before processing the next sample, ensuring that updates do not accumulate across the test set.

Algorithm Overview

We summarize our TTA method ASR-TRA in **Algorithm 1**, which shows how we integrate prompt intervention, counterfactual sampling, and reinforcement learning updates into a unified online "adaptation-and-prediction" loop.

Algorithm 1: **ASR-TRA**

Require: Input Mel-spectrogram s, Whisper model θ, CLAP reward model, number of samples n, prompt embedding p

Initialize Whisper model θ, set temperature $t = 0$
Generate deterministic output: $y_0 \leftarrow \text{Whisper}(s; \theta, t = 0)$
Compute reward: $r_0 \leftarrow \text{CLAP}(y_0, s)$
Insert random prompt p into Whisper decoder
for $i = 1, \ldots, n$ **do**
    Sample temperature $t_i \sim \text{Uniform}(0.4, 0.6)$
    Generate output with intervention: $y_i \leftarrow \text{Whisper}(s; \theta, \text{do}(p), t_i)$
    Compute reward: $r_i \leftarrow \text{CLAP}(y_i, s)$
end for
Compute advantages: $\text{adv}_i \leftarrow r_i - \text{mean}(\{r_j\}_{j=0}^{n})$
Compute policy gradient loss: $\mathcal{L} \leftarrow -\sum_i \text{adv}_i \cdot \log P_\theta(y_i|s, p)$
Update model parameters and prompt: $\theta \leftarrow \theta - \eta_1 \nabla_\theta \mathcal{L},\ p \leftarrow p - \eta_2 \nabla_p \mathcal{L}$
Generate adapted output: $y \leftarrow \text{Whisper}(s; \theta, p)$
Restore Whisper model θ and prompt p to original state
return Adapted transcription y

Experiments

Experimental Setup

We evaluate ASR-TRA on the lightweight Whisper-Tiny model (Radford et al. 2023), which contains approximately 39M parameters and is widely adopted in real-world ASR applications due to its low computational cost. Its compact size makes it well-suited for deployment in latency-sensitive or resource-constrained environments, such as on-device transcription or streaming ASR. However, Whisper-Tiny remains highly sensitive to distribution shifts, struggling to maintain performance under acoustic variations.

To enable fast and targeted adaptation, we insert a learnable decoder prompt of length four, introducing only $4 \times d = 1{,}536$ additional parameters, where $d = 384$ denotes the decoder embedding dimension. During adaptation, decoder temperatures are randomly sampled from a uniform range $[0.4, 0.6]$, with a total of 4 candidates generated in parallel by the Whisper-Tiny to encourage diverse hypotheses. Model parameters are updated using a learning rate η_1 in the range 10^{-6} to 10^{-5} depending on data complexity, while the prompt parameters are updated with a $100\times$ larger learning rate η_2, consistent with findings from prompt-tuning literature (Li and Liang 2021; Lester, Al-Rfou, and Constant 2021). All experiments are conducted on a single NVIDIA RTX 6000 Ada GPU.

Datasets and Evaluation Metrics

We consider two challenging distribution-shift scenarios that test robustness to both acoustic and linguistic variation:

1) Librispeech test-other with Environmental Noise. To evaluate robustness under acoustic corruption, we augment the LibriSpeech test-other set (Panayotov et al. 2015) with background noise sampled from the MS-SNSD corpus (Reddy et al. 2019). Eight additive noise types at 10 dB SNR are used: air conditioner, airport announcement, babble, copy machine, munching, neighbors, door shutting, and typing, following prior robustness benchmarks (Hendrycks and Dietterich 2019; Kim et al. 2023). Each utterance is augmented with one randomly sampled instance per noise type to approximate real-world environments.

2) L2-Arctic Non-Native Accented Speech. To assess sensitivity to speaker and pronunciation variation, we use the L2-Arctic dataset (Zhao et al. 2018), which contains English speech from speakers of six different first-language backgrounds. This condition introduces severe accent shifts that Whisper has not encountered during pretraining, thereby posing a strong challenge for evaluating its cross-speaker generalization ability.

These two settings together evaluate the adaptability of our method to acoustic domain shifts, a key challenge for real-world ASR deployment (Wang et al. 2021; Kim et al. 2023).

Evaluation Metrics. We report Word Error Rate (WER) for all experiments in this paper, following standard practice in test-time adaptation for ASR (Lin, Li, and Lee 2022; Kim et al. 2023). WER measures the minimum number of word-level insertions, deletions, and substitutions required to transform the model's transcription into the ground truth, normalized by the number of words in the reference transcript. Lower WER indicates better transcription accuracy. We also report latency in seconds to quantify the overhead introduced by the proposed TTA process.

Comparison to State-of-the-Art Approaches

We evaluate our method under two major types of distribution shift: environmental noise and speaker accent variation. Results are summarized in Tables 1 and 2, following prior work on test-time adaptation for ASR (Lin, Li, and Lee 2022; Kim et al. 2023; Wang et al. 2021).

Noise Robustness. We compare our approach with three baselines: (1) the original Whisper-Tiny model without adaptation (baseline), (2) SUTA (Lin, Li, and Lee 2022), and (3) SGEM (Kim et al. 2023). Table 1 reports the word error rate (WER) and inference latency on eight MS-SNSD noise conditions (Reddy et al. 2019) applied to LibriSpeech test-other (Panayotov et al. 2015). Our method achieves the lowest average WER (28.64%) and latency (0.720 s), consistently outperforming the baselines. Improvements are particularly pronounced under high-entropy noise such as *airport announcement* and *babble*, where acoustic variability is high. While SUTA marginally outperforms our method under the *neighbors* condition, our method remains significantly faster at inference time. This aligns with findings in

Dataset	Baseline WER (%) ↓	+ SUTA WER (%) ↓	+ SUTA Latency (s)	+ SGEM WER (%) ↓	+ SGEM Latency (s)	+ ASR-TRA (Ours) WER (%) ↓	+ ASR-TRA (Ours) Latency (s)
AC	26.54	24.02	1.683	23.76	0.730	**22.39**	**0.678**
AA	40.22	37.69	1.770	36.66	0.759	**35.66**	**0.670**
BA	36.85	40.12	1.643	33.92	0.710	**31.39**	**0.679**
CM	39.40	43.52	1.583	35.02	0.691	**34.72**	**0.631**
MU	29.98	34.12	1.645	34.40	0.705	**25.88**	**0.584**
NB	40.77	**33.75**	1.847	36.13	**0.756**	35.81	0.833
SD	23.12	22.91	1.656	**19.50**	0.909	22.34	**0.801**
TP	24.81	22.05	1.699	22.40	0.941	**20.91**	**0.886**
Mean	32.71	32.27	1.690	30.22	0.775	**28.64**	**0.720**

Table 1: WER (%) and inference latency (seconds) on Whisper-Tiny under eight noise conditions from MS-SNSD (SNR = 10 dB). Best values per row are in **bold**.

Setting	Baseline	SUTA	SGEM	ASR-TRA (Ours)
Arabic	32.74	35.94	35.19	**22.92**
Mandarin	28.03	27.68	28.03	**25.14**
Hindi	14.62	14.32	**12.61**	14.18
Korean	13.42	13.56	**12.12**	13.76
Spanish	42.64	41.25	41.78	**39.44**
Vietnamese	61.21	62.78	58.68	**53.84**
Mean	32.11	32.59	31.40	**28.21**

Table 2: WER (%) on L2-Arctic for English speech from speakers with different first-language (L1) backgrounds. Best result per row is in **bold**.

reward-guided adaptation literature, where additional diversity helps stabilize learning under heavy perturbations (Lee et al. 2025).

Accent Robustness. We further evaluate robustness to speaker and pronunciation variation using the L2-Arctic dataset (Zhao et al. 2018), which contains non-native English speakers from six different first-language (L1) backgrounds. Table 2 shows that our method achieves the best mean WER (28.21%), with substantial improvements on more challenging L1 groups such as Arabic and Vietnamese. These results suggest that our adaptation framework generalizes well across diverse phonetic systems, narrowing the gap between native and non-native speech recognition.

Overall, these findings demonstrate that our method enhances Whisper's robustness to both acoustic and linguistic distribution shifts while maintaining low inference latency. Compared with entropy-minimization-based methods (SUTA, SGEM), our reward-driven adaptation achieves a better balance between accuracy and efficiency, which is crucial for real-world ASR deployment.

Ablation Study

To better understand the contribution of each component, we conduct an ablation study on the LibriSpeech test-other dataset augmented with Gaussian noise at an SNR of 10 dB. In addition to Whisper-Tiny, we include Whisper-Base (74M vs. 39M parameters) to examine the effect of model size. It offers better recognition under noise while retaining efficient inference. We systematically vary three design dimensions in our adaptation framework:

- **Prompt Tuning**: Whether a learnable decoder prompt is injected during decoding.
- **Model Finetuning**: Whether model parameters are updated during test-time adaptation.
- **Reward Modeling**: Whether reward feedback is used, and whether it comes from CLAP (audio-text) or a pretrained LLM (text-text; we use DeepSeek V3 (Guo et al. 2024)), or a combination of both.

As shown in Table 3, all adaptation configurations outperform the unadapted Whisper baseline, highlighting the effectiveness of both prompt tuning and reward feedback. Several key observations emerge:

Effectiveness of Prompt Tuning. Comparing the second row (finetuning only, CLAP reward) and the fourth row (finetuning + prompt tuning, CLAP reward), we observe a notable WER reduction (40.49%→36.65% for Whisper-Tiny) accompanied by only marginal increases in inference latency. This suggests that the decoder prompt provides complementary adaptation capacity to parameter updates, improving alignment with reward signals.

Reward Modeling. Using CLAP as a reward model enables substantial WER improvements at negligible latency cost. Incorporating LLM-based feedback yields further gains (e.g., 0.3342→0.3024 for Whisper-Base) but adds a 7–9× latency overhead. The hybrid CLAP+LLM reward achieves the best overall WER. We also find that CLAP scores are negatively correlated with ground-truth WER (Spearman $\rho = -0.431$), supporting its reliability as a semantic reward.

Efficiency Considerations. We also explored LoRA-based parameter-efficient adaptation (Hu et al. 2021) but observed negligible speedup compared to full finetuning, as the majority of runtime is dominated by autoregressive decoding and reward computation. We therefore omit LoRA from further comparisons and instead focus on balancing reward complexity with inference efficiency.

Configuration			Whisper-Tiny		Whisper-Base	
Finetune	Prompt	Reward	WER(%) ↓	Latency (s)	WER(%) ↓	Latency (s)
N	N	N	45.06	–	40.25	–
Y	N	CLAP	40.49	0.486	35.97	0.573
N	Y	CLAP	42.75	0.472	37.45	0.552
Y	Y	CLAP	36.65	0.489	33.42	0.580
Y	Y	LLM	36.35	4.191	**30.24**	4.355
Y	Y	CLAP + LLM	**35.49**	4.365	30.28	4.367

Table 3: Ablation study of accuracy–latency trade-offs under different adaptation configurations on noisy LibriSpeech test-other. Each row enables a combination of prompt tuning, parameter finetuning, and reward feedback (CLAP, LLM, or both). WER (↓) and average inference latency (s/utterance) are reported for Whisper-Tiny and Whisper-Base.

Subset Evaluation on Confident Samples

We further analyze a subset of the LibriSpeech test-other set (Panayotov et al. 2015), comprising the 100 samples with the highest model confidence under additive Gaussian noise. Interestingly, the baseline WER on these high-confidence samples is as high as **83.61%**, which is *worse* than the WER on the full test set. This counterintuitive result indicates that Whisper-Tiny is often highly confident in its incorrect predictions, a phenomenon we refer to as *blind confidence*, where the model's internal certainty fails to reflect actual transcription accuracy under distribution shift.

As shown in Figure 2, SUTA (Lin, Li, and Lee 2022), which strongly depends on confidence-based entropy minimization, dramatically degrades performance to **122.37%**, exacerbating the misalignment between confidence and accuracy. SGEM (Kim et al. 2023), which incorporates sequence-level uncertainty, performs better at **63.00%**, but still fails to fully address blind confidence.

Our method ASR-TRA achieves the lowest WER of **45.17%** on this challenging subset, reducing errors by nearly half compared to the baseline. This robustness stems from the fact that our adaptation does not rely on model confidence or entropy as internal signals. Instead, it leverages external reward models (e.g., CLAP and LLM) to evaluate transcription quality, enabling it to revise predictions even when the model is falsely confident. By decoupling adaptation decisions from the model's own uncertainty estimates, our approach mitigates the effects of blind confidence.

These results reveal that internal confidence measures in Whisper-Tiny are not reliable under distribution shift, and that methods relying on confidence may be brittle when confronted with misleading inputs. Reward-guided adaptation provides a more reliable criterion, decoupling the adaptation signal from the model's own uncertainty estimates.

Discussion

Our experiments show that integrating causal reasoning with reinforcement learning enables effective test-time adaptation for Whisper-based ASR. Unlike confidence-driven baselines such as entropy minimization (Wang et al. 2021; Lin, Li, and Lee 2022) and pseudo-labeling (Lee 2013; Sohn et al. 2020), our causal intervention framework avoids error amplification from unreliable confidence signals and provides clearer interpretability, as confirmed by confidence-subset analysis where methods like SUTA tended to reinforce mistakes. Moreover, incorporating complementary external rewards further improves robustness to acoustic and linguistic shifts: CLAP offers fast and stable audio guidance, while DeepSeek V3 provides more precise semantic feedback at higher computational cost. Together, these components highlight the benefit of leveraging external cues rather than relying solely on internal confidence estimates for reliable adaptation under distribution shift.

Two limitations remain. First, our reward models have inherent constraints: LLM-based rewards dominate inference latency in hybrid settings, while CLAP currently supports mainly English audio–text similarity, limiting multilingual evaluation. Second, the framework focuses on single-utterance adaptation; extending to streaming or conversational ASR could enable persistent, context-aware robustness in real-world applications. Such a setting also offers the potential for implicit few-shot learning, where temporally accumulated feedback across utterances provides supervision akin to few-shot prompting.

Conclusion

We presented a causal reinforcement-learning framework, ASR-TRA, that adapts Whisper during inference through three lightweight stages. The adaptation begins by injecting a learnable decoder prompt as an explicit causal intervention on the decoding trajectory. To explore alternatives, the model generates diverse transcriptions via temperature-controlled sampling, which are then scored by CLAP to provide rewards. These rewards drive a policy-gradient update that fine-tunes the prompt and model parameters, gradually steering the model toward semantically improved outputs.

This loop consistently reduces WER on noisy and accented speech while introducing limited additional latency compared with existing test-time adaptation methods. By avoiding reliance on model-internal confidence and instead using an external, modality-aligned reward signal, the approach remains interpretable and robust under distribution shift. More broadly, framing test-time adaptation as a reward-driven causal process provides a promising direction for practical on-device or low-resource ASR, and suggests a natural path toward tighter integration of speech recognition with downstream multimodal or conversational systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62471420), Guangdong Basic and Applied Basic Research Foundation (2025A1515012296), and 2025 Tencent AI Lab Rhino-Bird Program.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33: 12449–12460.

Ben-David, S.; Blitzer, J.; Crammer, K.; Pereira, F.; et al. 2007. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 137–144.

Chan, W.; Jaitly, N.; Le, Q.; and Vinyals, O. 2016. Listen, Attend and Spell. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4960–4964.

Elizalde, B.; Kumar, P.; Zhan, J.; Shi, H.; et al. 2023. CLAP: Learning Audio-Text Joint Embeddings from Large-Scale Weakly-Supervised Data. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Guo, Z.; Li, X.; Wang, H.; Zhang, L.; Liu, R.; Zhao, Z.; Wu, Y.; et al. 2024. DeepSeek-V3: Scaling Open-Source Language Models with Mixture-of-Experts and Reinforcement Learning. *arXiv preprint arXiv:2401.06066*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations (ICLR)*.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations (ICLR)*.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, L.; Wang, L.; and Li, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Kim, C.; Park, J.; Shim, H.; and Yang, E. 2023. SGEM: Test-Time Adaptation for Automatic Speech Recognition via Sequential-Level Generalized Entropy Minimization. *arXiv preprint arXiv:2306.01981*.

Ko, T.; Peddinti, V.; Povey, D.; and Khudanpur, S. 2015. Audio augmentation for speech recognition. In *Proceedings of Interspeech*, 3586–3589.

Lee, D.-H. 2013. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *ICML Workshop on Challenges in Representation Learning*.

Lee, T.; Chottananurak, S.; Kim, J.; Shin, J.; Gong, T.; and Lee, S. 2025. Test-Time Adaptation with Binary Feedback. *arXiv preprint arXiv:2505.18514*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale: Parameter-Efficient Adaptation for Pretrained Language Models. In *Proceedings of EMNLP*.

Levine, S.; Finn, C.; Darrell, T.; and Abbeel, P. 2016. End-to-end training of deep visuomotor policies. In *Journal of Machine Learning Research*, 1, 1334–1373.

Li, J.; Deng, L.; and Gong, Y. 2020. Multi-condition training for robust automatic speech recognition. In *Robust Automatic Speech Recognition: A Bridge to Practical Applications*, 43–72. Academic Press.

Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky, D. 2016. Deep reinforcement learning for dialogue generation. In *EMNLP*, 1192–1202.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of ACL*.

Lin, G.-T.; Li, S.-W.; and Lee, H.-y. 2022. Listen, Adapt, Better WER: Source-free Single-Utterance Test-Time Adaptation for Automatic Speech Recognition. *arXiv preprint arXiv:2203.14222*.

Logan, B. 2000. Mel Frequency Cepstral Coefficients for Speech and Speaker Recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

McFee, B.; Raffel, C.; Liang, D. P.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proc. 14th Python in Science Conference*, 18–25.

Mitrovic, J.; McWilliams, B.; Walker, J.; Buesing, L.; and Blundell, C. 2021. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations (ICLR)*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Radford, A.; Kim, J. W.; Xu, T.; Jeong, G.; et al. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint arXiv:2212.04356*.

Reddy, C. K.; Dubey, H.; Garg, V.; Cheng, R.; Cutler, R.; and Gehrke, J. 2019. The Microsoft Scalable Noisy Speech Dataset (MS-SNSD). Technical report, Microsoft. Microsoft Technical Report.

Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised Pre-training for Speech Recognition. In *Interspeech*, 3465–3469.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; et al. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems*, 33: 596–608.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press, 2 edition.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, D.; Zhu, E.; Yue, X.; and Gonzalez, J. E. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations (ICLR)*.

Williams, R. J. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. In *Machine Learning*, 3–4, 229–256.

Zhao, G.; Hirst, D.; Povey, D.; and Khudanpur, S. 2018. L2-ARCTIC: A Non-Native English Speech Corpus. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5204–5208.

Zhao, S.; Wang, X.; Zhu, L.; and Yang, Y. 2024. Test-Time Adaptation with CLIP Reward for Zero-Shot Generalization in Vision-Language Models. In *International Conference on Learning Representations (ICLR)*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T.; et al. 2019. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*.