When Reward Doesn't See Everything: Modality-Blind Credit Assignment for Simultaneous Speech-to-Speech Translation

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning is increasingly used to fine-tune simultaneous speech-to-speech translation (S2ST) models. These models produce both text and audio at the same time, but the reward functions used to train them are computed almost entirely from text. We show that this mismatch causes a failure the training process cannot detect on its own, in which the training reward rises steadily while the model's own voice quality collapses beneath it. We test this by applying group relative policy optimization (GRPO) with a text-only bilingual evaluation understudy (BLEU) reward to Hibiki-Zero, a 3-billion-parameter S2ST model. During training, the reward score rises steadily, but translation quality on held-out data does not follow it, and audio quality falls sharply. Speaker similarity drops by more than 40%, and naturalness, measured by the UTokyo-SaruLab mean opinion score (UTMOS), falls from 3.410 to 2.610. We trace this to GRPO's credit assignment, which broadcasts a single scalar advantage derived from the text reward across every token the model produces, including the audio tokens the reward never examines. This is not simply a missing safeguard: the collapse persists even when we restore GRPO's Kullback-Leibler (KL) divergence penalty, which limits how far the model may drift from its starting point. Holding the reward fixed, we substitute a learned value function that scores each position separately, using proximal policy optimization with generalized advantage estimation (PPO+GAE). This removes the collapse entirely, and it holds when group size is matched to GRPO's. We further test four cheaper alternatives, including one with no learned value estimator at all. All four collapse and are statistically indistinguishable from GRPO. This shows that what matters is not whether a value estimator is present, but whether credit is assigned by comparing samples

1 Introduction

Simultaneous speech-to-speech translation (S2ST) requires a model to continuously trade translation quality against real-time latency. Reinforcement learning (RL) has become the standard tool for tuning this trade-off. Recent S2ST systems apply RL algorithms such as group relative policy optimization (GRPO; Shao et al., 2024) and proximal policy optimization (PPO; Schulman et al., 2017) to fine-tune policies that jointly and synchronously produce two output streams, a text stream and an audio-codebook stream. In every such system to date, the reward driving this training is computed entirely from the text stream.However, the audio stream the same policy produces is never evaluated by the reward at all. This asymmetry is not harmful on its own, since a text-derived reward can still correlate with overall output quality. It leaves open a basic question: can the credit-assignment mechanism that turns this reward into per-token updates distinguish text tokens from audio tokens it was never designed to grade, and if it cannot, what happens to the stream the reward cannot see?

Existing RL-tuned S2ST systems, including Hibiki-Zero (Labiausse et al., 2026) and Seed-LiveInterpret 2.0 (Cheng et al., 2025), do not evaluate audio quality independently of the training reward, so a failure confined to the audio stream would not surface in their own reported metrics. We close this gap directly, and show that a purely text-derived reward can raise the training signal while audio quality collapses beneath it, undetected by the objective being optimized. The mechanism we identify is not specific to speech. Wherever a single scalar reward supervises a policy that emits

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

several coupled output streams, the same structural gap between what the reward observes and what the update reaches can arise. Analogous reward-proxy failures are documented in text-to-speech synthesis, where a transcription-oriented reward degrades prosody (Shin et al., 2025; Wang et al., 2026), and in text reinforcement learning from human feedback, where a spurious length correlation is corrected at the reward-model level (Chen et al., 2024). Our results point the other way in the multistream setting, locating the correctable component in credit assignment rather than in the reward.

This paper makes the following contributions.

1. We identify a previously undocumented failure mode in RL-trained simultaneous S2ST. A text-only reward optimized with GRPO raises the training reward while degrading speaker identity and naturalness. We demonstrate this failure on a released Hibiki-Zero model and show that it reproduces under an independent random seed and a second language pair.
2. We trace the failure to a structural weakness in critic-free multistream RL, which we call cross-modal advantage broadcasting. GRPO computes one scalar advantage from the text reward and applies it to synchronized text and audio-codebook streams, including audio tokens that the reward never evaluates. The collapse persists with a KL penalty matched to a full-critic baseline, ruling out missing regularization as the cause. Testing this mechanism directly through a dose-response ablation, varying the weight of the text-derived advantage on the audio stream reveals non-monotonic dynamics, indicating that the collapse mechanism extends beyond uniform advantage broadcasting alone.
3. We show that a full value function eliminates the collapse. Specifically, PPO with GAE prevents degradation while using the same text-only reward and group size as GRPO, isolating the cause to how credit is assigned rather than to what the reward measures. We further show that this fix generalizes to a second language pair.
4. We evaluate four lightweight approximations to a full critic, including a critic-free segment-level baseline, and find that none prevents the collapse. All four remain statistically indistinguishable from GRPO, showing that position-specific value estimation rather than group-relative comparison is decisive.

2 Related Work

Early S2ST systems relied on cascaded automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) pipelines, which introduce compounding latency and lose speaker identity and prosody at each hand-off. End-to-end systems address this limitation directly, though each leaves open the specific question this paper investigates. Table 1 summarizes these systems, their contribution, and the limitation this paper addresses. The gap it identifies is that Hibiki-Zero's RL stage evaluates its own success using the same text-derived reward it optimizes (Labiausse et al., 2026), therefore a failure confined to the audio stream it also produces would not surface in its own reported metrics. Although Seed-LiveInterpret 2.0 attributes training instability and reward hacking to interacting reward terms (Cheng et al., 2025), its ablation studies do not determine whether the issue lies in reward design or in token-level credit assignment. We address this question directly, using Seed-LiveInterpret 2.0's reward structure as the basis for our PPO with GAE baseline.

GRPO replaces PPO's learned value function with a group-normalized reward baseline, and is the algorithm Hibiki-Zero adapts to its multistream setting (Shao et al., 2024). Recent large-scale RL systems including Magistral omit its KL-divergence term to reduce memory overhead, a design choice Hibiki-Zero follows and one we test directly (Rastogi et al., 2025). This omission forgoes a per-token, per-stream baseline, which PPO with GAE instead provides through a position-specific value function at the cost of a separate critic network (Schulman et al., 2017; 2015). VinePPO and Segment Policy Optimization pursue cheaper alternatives to a full critic at finer granularity than GRPO's trajectory-level advantage, reporting that finer-grained critic-free credit assignment can match a full critic on text-generation reasoning tasks (Kazemnejad et al., 2024; Guo et al., 2025). We test this direction directly with a segment-level, critic-free baseline for the audio stream, and find that it collapses at the same statistical level as our critic-based lightweight alternatives, indicating that the gap these methods close in single-stream text tasks does not close analogously once credit must be split across two output modalities. However, prior comparisons do not hold

Table 1: Prior S2ST and RL credit-assignment systems, their contribution, and their limitations.

System	Contribution	Limitations
Seamless (Barrault et al., 2023)	Streaming multilingual translation via monotonic attention.	No RL stage; latency-quality trade-off fixed post-training.
StreamSpeech (Zhang et al., 2024)	Joint read/write and translation learning.	No separate RL stage to test reward-driven credit assignment.
Hibiki (Labiausse et al., 2025)	Multistream RQ-Transformer architecture.	Requires hand-annotated word-level alignments.
Hibiki-Zero (Labiausse et al., 2026)	Removes alignment dependency; GRPO with text-only reward.	Evaluates success via the same text-only signal it optimizes.
Seed-LiveInterpret 2.0 (Cheng et al., 2025)	PPO with stepwise and sequence-level reward for duplex S2S.	Reports instability but does not isolate reward from credit assignment.
VinePPO (Kazemnejad et al., 2024) / SPO (Guo et al., 2025)	Finer-grained critic-free credit assignment for text tasks.	Tested as a critic-free baseline in this work; fails at the same statistical level as our critic-based alternatives.

the reward formulation fixed while changing only the credit-assignment mechanism. Our matched-reward comparison in Section 4 performs this isolation directly.

Several systems address simultaneity without an RL stage that could exhibit the failure studied in this paper. Seamless (Barrault et al., 2023) and StreamSpeech (Zhang et al., 2024) improve streaming S2ST through architectural design and multi-task learning, rather than RL fine-tuning. As a result, they do not expose the reward-driven credit-assignment mechanism we evaluate. Hibiki (Labiausse et al., 2025), the supervised predecessor of Hibiki-Zero, relies on hand-annotated word-level alignments and does not use RL. Hibiki-Zero removes this alignment requirement through GRPO, which is precisely the training stage our work examines.

Other recent methods apply RL-based credit assignment to simultaneous translation, but they operate on text-only or speech-to-text outputs. SeqPO-SiMT (Xu et al., 2025), SimulPL (Yu et al., 2025), and Ouyang et al.'s hierarchical policy optimization (Ouyang et al., 2026) each optimize a single text-derived reward stream through sequential policy optimization, preference alignment, or hierarchical optimization. None of these systems jointly generates a second output modality that remains unobserved by the reward. Thus, they do not test the audio-specific credit-assignment failure isolated in this work.

The failure this paper characterizes, a reward computed over one output dimension degrading another it never measures, has documented analogues outside S2ST. In text-to-speech synthesis, GRPO trained on a transcription-oriented reward collapses prosody into monotone speech even as intelligibility improves. This instability recurs across a broader family of GRPO-based TTS systems balancing multiple acoustic objectives (Shin et al., 2025; Wang et al., 2026). This is the closest direct analogue to our finding, but neither line of work isolates whether the cause is the reward's incompleteness or the credit-assignment mechanism propagating it, the distinction our matched-reward experiment is built to separate. A comparable case of reward proxy divergence appears in text reinforcement learning from human feedback, where a spurious correlation between response length and reward is addressed by disentangling the two signals at the reward-model level, a fix that presumes the reward itself is the correctable component (Chen et al., 2024). The same pathology has recently been identified in joint audio-video generation, where a single global advantage over multi-modal diffusion output is similarly shown to be insufficient, motivating per-modality advantage routing as a remedy (Zhang et al., 2026). Our results point the other way in S2ST's multistream setting: replacing GRPO's credit assignment with a full value function eliminates the collapse using the same text-only reward, isolating the failure to how credit reaches the unobserved stream rather than to what the reward measures. To our knowledge, no prior work documents this failure mode in simultaneous S2ST or isolates it to this mechanism specifically.

3 The Proposed Methodology

Our method tests whether GRPO causes audio collapse through its credit-assignment mechanism, rather than through the text-only reward itself. To isolate this effect, we keep the model architecture, training data, reward definition, and compute budget fixed across comparisons whenever possible,

and state the cases where complete isolation is not feasible, since GRPO and critic-based methods differ in their optimization structure. The evaluation proceeds in three stages: first, we define the model, dataset, reward, and held-out evaluation protocol that remain fixed across all experiments (Section 3.1); second, we compare seven RL configurations, each designed to test a specific explanation for the observed collapse or a candidate mechanism for preventing it (Section 3.2); third, we run five diagnostic experiments that examine the stability of our conclusions across reward matching, language pairs, random seeds, and evaluation-set choices (Section 3.3).

3.1 Model, Data, and Evaluation Protocol

We evaluate Hibiki-Zero (Labiausse et al., 2026), a 3-billion-parameter multistream Transformer, using the released `hibiki-zero-3b` checkpoint for French-to-English simultaneous speech-to-speech translation. We use a real, released system, rather than training from scratch, so that any collapse we observe reflects an active property of deployed S2ST training rather than a failure constructed to fit our hypothesis. The model jointly emits a text token y_t^{text} and $K = 16$ discrete audio-codebook channels $y_t^{a,1:K}$ via the Mimi neural audio codec at 12.5 Hz, referred to as the text and audio streams. Baseline (no RL) refers to this checkpoint with no fine-tuning applied.

All fine-tuning uses low-rank adaptation (LoRA, rank 16) with 4-bit NormalFloat (NF4) quantization, at a learning rate of 1×10^{-5}, fifty times Hibiki-Zero's reported rate, since the original rate produced no measurable learning at our batch scale. This configuration is held fixed across every variant so that it cannot itself become a confound (full details in Appendix A). RL fine-tuning uses the Europarl-ST v1.1 French-English train split, filtered to 2–20s utterances to ensure enough tokens per utterance for reward checkpointing. We train on Europarl-ST alone, rather than Hibiki-Zero's own Europarl-ST/CoVoST2 mixture, so that any collapse can be attributed to the training procedure rather than to heterogeneous data quality.

The training reward is computed entirely from the text stream, matching how Hibiki-Zero is actually trained, since this asymmetry is what our paper investigates. We approximate Hibiki-Zero's word-level reward checkpointing, placing $s = \max(1, \text{round}(|\text{source}|/n_w))$ checkpoints at proportional output positions, since the true source-word boundary per output frame is not exposed by the model. Each checkpoint mixes partial and final-sequence BLEU,

$$r_t = \alpha \cdot \text{BLEU}_{\text{inter}}(t) + (1 - \alpha) \cdot \text{BLEU}_{\text{final}}, \tag{1}$$

with $\alpha = 0.4$, matching Hibiki-Zero's own reported mixing weight, so that partial-hypothesis progress and final-sequence quality both contribute to the reward at each checkpoint. As no training code for Hibiki-Zero's RL stage is public, this reconstruction is our own interpretation of the published equations, which we note as a limitation. Figure 1 summarizes how the seven variants in Section 3.2 route credit from this reward to the two streams.

We evaluate every checkpoint on a fixed, held-out set of 150 Europarl-ST utterances, kept separate from training and, apart from one diagnostic that tests the effect of reuse (Section 3.3), never used to select which checkpoint to report. We measure translation quality with ASR-BLEU (Papineni et al., 2002), naturalness with UTMOS22 (Saeki et al., 2022), speaker identity with ECAPA-TDNN (Desplanques et al., 2020) cosine similarity, and simultaneity with LAAL (Papi et al., 2022). This set is necessary because a collapse confined to the audio stream is only detectable by measuring that stream independently of the text-derived training signal. Every comparison uses paired bootstrap resampling (10,000 resamples) and the Wilcoxon signed-rank test, Bonferroni-corrected within each family, requiring both tests to agree given BLEU-family metrics' non-normality.

3.2 Reinforcement Learning Variants

We compare seven RL configurations, each answering a specific question about where GRPO's failure originates and what resolves it. Figure 1 summarizes how each assigns credit from the reward to the two output streams.

Vanilla GRPO reproduces Hibiki-Zero's own released training recipe unmodified, so any collapse we observe reflects an active property of deployed S2ST training rather than a hypothetical concern constructed to fit our argument.

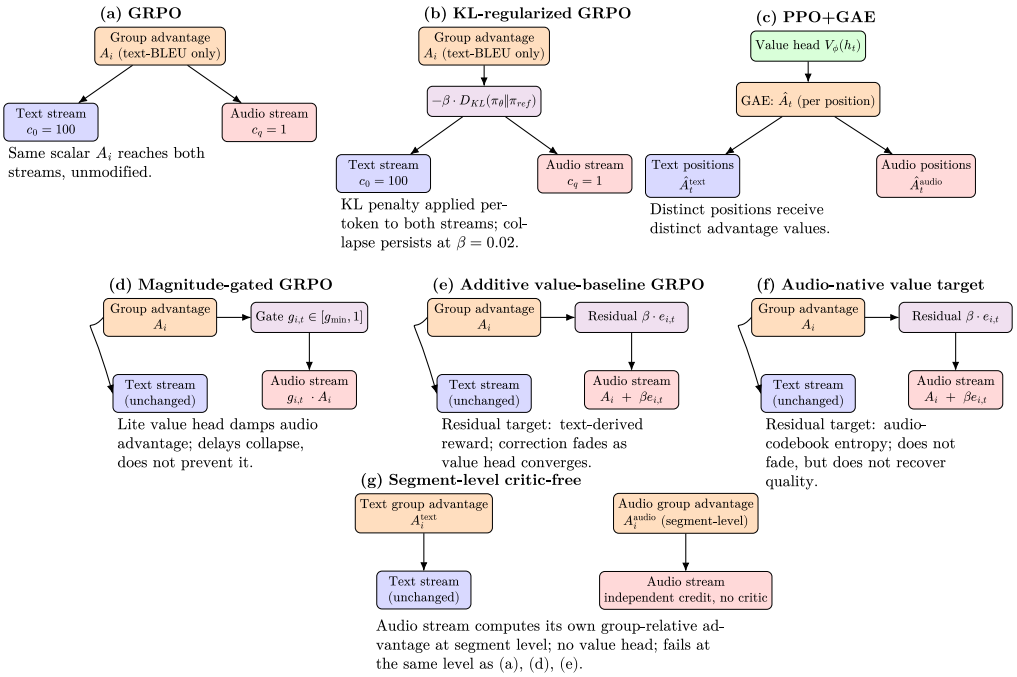


Figure 1: **Credit assignment across all seven RL variants.** GRPO (a) broadcasts one unmodified advantage to both streams; (b) adds a KL penalty without preventing the collapse; PPO+GAE (c) instead computes a distinct, position-specific advantage per stream. Panels (d–f) test three lightweight approximations to a full value function, and (g) tests a critic-free alternative in which the audio stream computes its own independent, segment-level advantage. Section 4 reports that only (c) avoids the collapse.

KL-regularized GRPO adds a KL-divergence penalty against the frozen reference policy, matched to our PPO+GAE baseline ($\beta = 0.02$). Hibiki-Zero's own recipe omits this penalty to save memory, so we test whether simply restoring it resolves the collapse before attributing the failure to anything more structural.

PPO with Generalized Advantage Estimation (PPO+GAE) replaces GRPO's group-relative advantage with a learned, position-specific value function, following the general reward structure of Seed-LiveInterpret 2.0 (Cheng et al., 2025). This is our positive control: if a per-token, per-stream baseline prevents the collapse, GRPO's failure is attributable to credit assignment rather than to RL fine-tuning generally. GRPO and PPO+GAE also differ in group size, however, which could independently explain a stability difference; we close this second confound directly in Section 4.

Three lightweight critic approximations (a magnitude gate, an additive value-baseline correction, and an audio-native value target) test whether a full critic's benefit can be recovered more cheaply, using a value head 0.20% its size (2,049 vs. 1,049,601 parameters), since a full critic is materially more expensive to train. Each modifies only the audio stream's advantage, isolating whether architecture, regression target, or correction fade rate determines success or failure.

A segment-level, critic-free baseline removes the value function entirely: the audio stream computes its own independent, segment-level group-relative advantage, rather than a copy or correction of the text-derived advantage, while the text stream's GRPO update is left unchanged.

Every hyperparameter not specific to a variant's own mechanism is held identical across all seven configurations; full values are reported in Appendix A.

3.3 Robustness Diagnostics

Beyond the seven primary variants, we conduct five diagnostic experiments to test the stability of our central claims. These experiments examine whether GRPO collapses because of credit assignment rather than reward content, and whether the stability of PPO with GAE reflects the method itself rather than our experimental choices.

- **Matched-reward PPO+GAE.** We isolate a confound in the original GRPO/PPO+GAE comparison: the two methods differ in reward formulation as well as credit-assignment mechanism, since PPO+GAE's reward includes a richer multi-turn component GRPO's does not. We retrain PPO+GAE using GRPO's exact text-only reward, with every other hyperparameter unchanged, testing whether PPO+GAE's stability depends on its richer reward or on its credit-assignment mechanism alone.
- **Cross-modal advantage dose-response.** We test our named mechanism directly. If cross-modal advantage broadcasting causes the collapse, scaling c_q, the weight applied to the text-derived advantage on the audio stream, should scale the resulting damage. We train GRPO at four values of $c_q \in \{0, 0.1, 0.5, 1\}$, holding every other hyperparameter fixed, and evaluate whether audio quality degrades monotonically as c_q increases.
- **A second, disjoint held-out set.** We test whether our results depend on reusing the same set for checkpoint selection and final evaluation. We build an independent 150-utterance evaluation set and re-evaluate every reported checkpoint on it without retraining, checking whether the same pattern holds on data that played no role in selecting which checkpoints to report.
- **PPO+GAE on a second language pair.** We extend our generalization check beyond GRPO's own collapse, which we confirm generalizes to Spanish-English (Section 4), testing whether PPO+GAE's fix generalizes as well, or is specific to our primary language pair.
- **PPO+GAE at matched group size.** We isolate a second confound between GRPO and PPO+GAE beyond reward formulation: GRPO samples multiple rollouts per prompt to compute its group-relative advantage, while PPO+GAE as described above uses only one. We re-run PPO+GAE at GRPO's own group size, with every other hyperparameter unchanged, testing whether PPO+GAE's stability depends on this difference in sampling regime or on its credit-assignment mechanism alone.

Full configuration for all five diagnostics matches the corresponding primary variant (Table 4, Appendix A), except where a diagnostic's own manipulation requires a difference, noted above.

4 Results and Analysis

We begin by showing that vanilla group relative policy optimization reproduces the audio collapse that motivates this work, then use the remaining variants and diagnostic checks to isolate its cause. Table 2 compares all seven primary RL variants against the baseline on held-out data. GRPO's own reported checkpoint is already a significant regression from baseline on speaker similarity and LAAL, the two metrics that most reliably distinguish a stable run from a collapsing one, and the collapse deepens steadily rather than plateauing when held past that point, reaching a 40.6% relative drop in speaker similarity by step 1000. Figure 2(a) shows this against the rising training reward: the signal driving training and the quality a listener would actually hear move in opposite directions. KL-regularized GRPO does not rescue this. It regresses on ASR-BLEU, UTMOS, speaker similarity, and LAAL relative to baseline, performing no better than vanilla GRPO on ASR-BLEU despite the added penalty, and is not statistically distinguishable from unregularized GRPO's own step-600 checkpoint on UTMOS ($p_{\text{boot}} = 0.68$), the metric the penalty was intended to protect; its one benefit is a measurable reduction in fully degenerate outputs, a robustness gain distinct from average-case quality.

Measuring the model's own text stream directly, rather than transcribing its audio output first, tells a related but distinct story. Text-BLEU exceeds ASR-BLEU on every checkpoint, by 0.3 to 4.6 points, with the largest gap on the audio-native value target, the most audio-degraded variant in Table 2 (24.22 vs 28.84). Since ASR-BLEU depends on successfully transcribing degraded audio, this gap is consistent with part of the collapse being an audio-synthesis problem that confounds transcription, rather than a loss of the underlying translation competence itself.

PPO+GAE is the only variant statistically indistinguishable from baseline on every metric. To confirm this stability comes from how it assigns credit rather than what it is rewarded for, we retrain it using GRPO's own text-only reward in place of its richer one; the result sits within noise of the original run on every metric (all differences below 0.01 speaker-similarity and 0.01 UTMOS

Table 2: Held-out evaluation of the baseline and all seven RL variants, at each variant's reported checkpoint. Text-BLEU scores the model's own text stream directly; ASR-BLEU transcribes its audio output first. Arrows indicate whether higher (↑) or lower (↓) values are better. † marks a statistically significant regression relative to baseline (paired bootstrap and Wilcoxon signed-rank test, both significant after Bonferroni correction); full confidence intervals are reported in Appendix B.

Method	Checkpoint	ASR-BLEU↑	Text-BLEU↑	UTMOS↑	Speaker Sim.↑	LAAL (s)↓
Baseline (no RL)	—	26.67	27.39	3.410	0.4531	2.790
GRPO	step 300	25.53	28.04	3.372	0.4277†	3.095†
KL-Regularized GRPO	step 600	25.16	27.81	3.024†	0.3745†	3.460†
PPO+GAE	step 600	27.75	28.01	3.410	0.4512	2.733
Matched-Reward PPO+GAE	step 600	26.44	26.79	3.400	0.4508	2.721
Magnitude-Gated GRPO	step 600	25.13	27.35	3.032†	0.3672†	3.438†
Additive Value-Baseline GRPO	step 600	25.48	28.74	3.100†	0.3918†	3.409†
Audio-Native Value Target	step 600	24.22	28.84	2.911	0.3345	3.485
Segment-Level Critic-Free	step 600	27.66	29.99	3.103†	0.3713†	3.487†

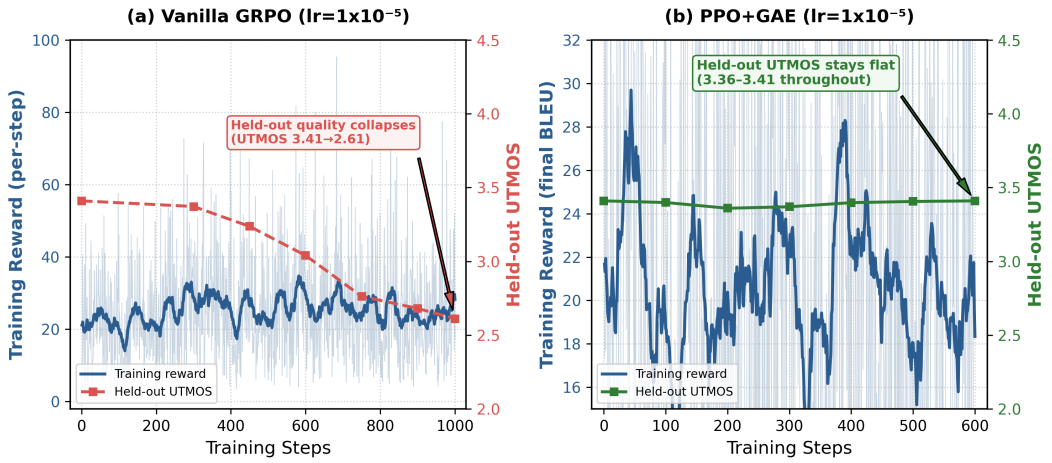


Figure 2: A text-only training reward can rise while held-out audio quality falls beneath it. (a) Vanilla GRPO: training reward (blue, left axis, 25-step rolling mean) rises across 1000 steps while held-out UTMOS (red, right axis) falls from 3.41 to 2.61. (b) PPO+GAE, same axes and 600-step span: both stay flat and close to baseline.

points). This matched-reward comparison is the paper's central result: the only thing that changed between GRPO and matched-reward PPO+GAE is how each turns reward into a per-token update, and PPO+GAE alone avoids the collapse, so the cause must lie in credit assignment rather than in what the reward measures. The remaining four lightweight alternatives to a full value function all regress from baseline on the metrics defining the collapse; three modify the audio stream's advantage through a small value head, while the fourth removes the value function entirely and gives the audio stream its own independent, segment-level advantage. All four remain statistically indistinguishable from vanilla GRPO, indicating the failure is not specific to having a learned critic but to the coarser, group-relative way credit is computed in the first place.

Figure 3 shows why this matters beyond the final checkpoint: the lightweight alternatives track baseline closely through roughly step 300 and only diverge afterward, so a snapshot at that step alone would have suggested each was working when the collapse was merely delayed. Together, these results answer the question posed at the outset: GRPO's credit-assignment mechanism cannot distinguish text tokens from audio tokens the reward never observes, and the unobserved stream absorbs a biased update as a consequence.

Three diagnostics test whether this conclusion survives our own experimental choices. GRPO and PPO+GAE differ not only in credit assignment but in group size, since GRPO samples four rollouts per prompt and PPO+GAE above samples one; re-running PPO+GAE at GRPO's own group size yields results statistically indistinguishable from both baseline and the single-rollout run on all four metrics (ASR-BLEU 26.55, UTMOS 3.408, speaker similarity 0.4529, LAAL 2.70s), while remaining significantly better than GRPO on UTMOS, speaker similarity, and LAAL after correction, closing this confound. The core collapse reproduces under an independent random seed and gener-

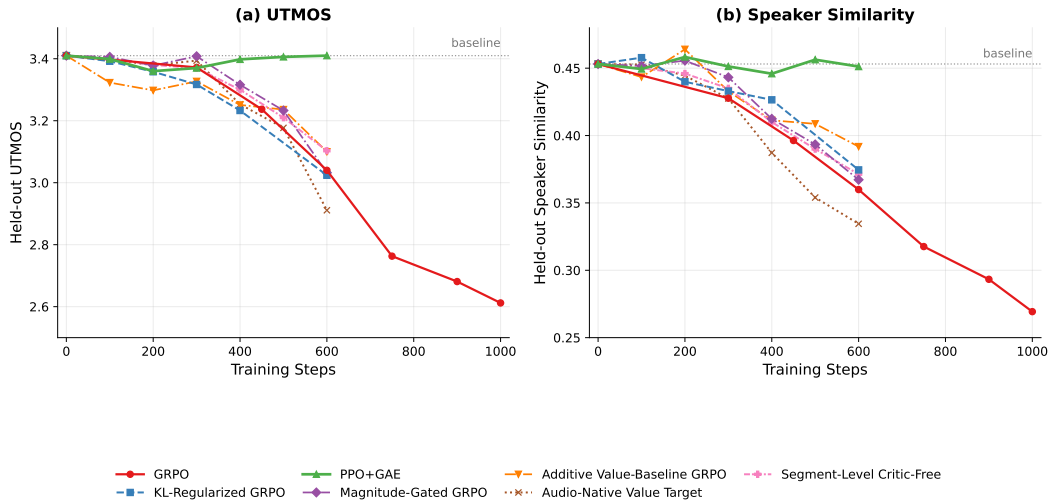


Figure 3: Held-out audio quality across training, for all GRPO-family variants and PPO+GAE. (a) UTMOS and (b) speaker similarity against training step, untrained baseline as a dashed line. PPO+GAE (green) is the only trajectory remaining within the baseline band through step 600; every other variant, including the segment-level critic-free baseline, diverges downward after step 300.

alizes, along with PPO+GAE's fix, to Spanish-English (Google FLEURS), where speaker similarity trends the same direction without reaching significance. Finally, because every checkpoint above was selected on the same held-out set used for reporting, we re-evaluated all of them on a disjoint 150-utterance set without retraining: PPO+GAE again shows no regression from baseline and again beats GRPO on LAAL, though some effect sizes shrink and one originally-significant result does not survive, indicating our magnitudes were somewhat inflated by selecting on the reported set without overturning which method wins or why.

Finally, we test the mechanism we name in this paper, cross-modal advantage broadcasting, directly (Table 3). Under this account the audio stream degrades because it receives the same text-derived advantage as the text stream, scaled by a fixed weight c_q; if correct, $c_q = 0$ should fully protect audio quality while $c_q = 1$ should reproduce the full collapse. It does not. At step 600, $c_q = 0$ is tied for the worst point on both UTMOS and speaker similarity, an ordering already visible but not significant at step 300. This is the opposite of what a simple linear-leakage account predicts. We tested the most direct explanation, that the shared LoRA backbone continues drifting under the unopposed text-only loss at $c_q = 0$ with no audio-specific counter-signal; weight-norm drift rules this out, since it scales roughly monotonically with c_q (0.864 at $c_q = 0$ rising to 1.174 at $c_q = 1$), the opposite clustering from collapse severity. How far the weights moved is therefore not what determines how far audio quality fell, and we leave identifying the true driver to future work (Section 5). Credit assignment remains the established cause of the collapse through the matched-reward comparison; what this test complicates is only the specific account of *why* the unobserved stream absorbs a biased update, not that it does.

Table 3: Cross-modal advantage dose-response at step 600. c_q is the weight applied to the text-derived advantage on the audio stream; $c_q = 1$ reproduces vanilla GRPO. Baseline: ASR-BLEU 26.67, UTMOS 3.410, Speaker Sim. 0.4531, LAAL 2.790s. $c_q = 1$ significantly outperforms $c_q = 0$ on UTMOS ($p_{\text{boot}} < 0.0001$, $p_{\text{Wilcoxon}} < 0.0001$), the opposite of what a simple linear-leakage account predicts.

c_q	ASR-BLEU↑	UTMOS↑	Speaker Sim.↑	LAAL (s)↓
0	24.06	2.749	0.3461	3.348
0.1	25.57	2.792	0.3464	3.332
0.5	25.40	3.084	0.3816	3.373
1	24.36	3.040	0.3599	3.452

5 DISCUSSION

Limitations. Our experiments use LoRA and 4-bit quantization on a single consumer GPU, evaluating one architecture (Hibiki-Zero) on two language pairs, so our results characterize this failure mode in one system rather than the full space of S2ST models. We measure audio quality with UTMOS and speaker similarity, following prior work, but a human listening study would further confirm these findings. Our reward is reconstructed from Hibiki-Zero's published equations, since no training code is public, and we could only sensitivity-check this reconstruction for the GRPO-family variants; PPO+GAE's intermediate reward values were not logged during training. The c_q dose-response test also complicates our named mechanism: blocking the audio stream's advantage entirely ($c_q = 0$) does not protect audio quality, and weight-norm data rules out raw parameter drift as the explanation without offering a confirmed one in its place. Credit assignment remains the established cause of the collapse through our matched-reward comparison; only the precise way c_q affects severity stays open.

Future work. The mechanism we identify, group-relative credit assignment broadcasting a scalar advantage across streams a reward does not jointly observe, is not specific to text-audio S2ST. An important direction is testing whether it generalizes to other multistream policies trained with critic-free RL, such as joint video-audio generation or tool-use agents emitting both reasoning and structured actions under a single outcome reward. A second direction is theoretical: characterizing, for a general class of group-relative advantage estimators, the conditions under which an unobserved output stream is guaranteed a biased update, so practitioners could predict this failure analytically rather than discover it per architecture. Our four lightweight alternatives, including one with no critic at all, suggest the open problem is not a cheaper critic, but any per-stream signal, critic-based or not, that is both audio-native and reward-correlated.

6 CONCLUSION

This paper identifies a previously undocumented failure mode in reinforcement learning (RL)-trained simultaneous speech-to-speech translation (S2ST): a model can improve its text-based training reward while its generated speech becomes substantially worse. We demonstrate this on Hibiki-Zero, where GRPO with a text-only reward increases the training signal, yet held-out speaker similarity drops by more than 40%, and naturalness decreases from 3.410 to 2.610. We trace this divergence to cross-modal advantage broadcasting, where one group-relative advantage is applied uniformly to both text and audio-codebook tokens, including audio tokens the reward never evaluates. The collapse persists under KL regularization and reproduces under an independent seed and a second language pair, ruling out a simple tuning artifact. A matched-reward comparison then isolates the cause: PPO+GAE eliminates the collapse using the same text-only reward as GRPO, and the result holds when group size is also matched, closing every confound between them. Four lower-cost alternatives to a full value function fail to match this stability, including one that assigns credit independently per stream with no learned critic at all, indicating the failure depends on the granularity of credit assignment rather than on the presence of a critic specifically. A direct test of the broadcasting mechanism yields a non-monotonic result, showing its dynamics are more complex than uniform broadcasting alone predicts, even as the matched-reward comparison establishes credit assignment as the underlying cause. These results show that multistream generation under partially observed rewards requires per-token, per-stream credit assignment, not scalar critic-free updates.

AI USAGE DISCLOSURE

Generative AI tools like Grammarly and Gemini were employed solely as editing assistants for language polish, sentence restructuring, and LaTeX formatting. No core scientific content, claims, experiments, or figures were generated by AI, and all text was thoroughly verified and edited by the authors.

REFERENCES

Loïc Barrault, Yu-An Chen, Paul-Ambroise Duquenne, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, John Hoffman, Christopher Klaiber, Peng Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Peloquin, Kaushik Ramadan, Abinesh Ramesh, Anna Sun, Bokai Wang, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Yin, Pierre Andrews, Ankur Bapna, Christophe Baume, Maha Fazel-Zarandi, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023. URL https://arxiv.org/abs/2312.05187.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*, 2024. URL https://arxiv.org/abs/2402.07319.

Shanbo Cheng, Yu Bao, Zhichao Huang, Yu Lu, Nan Peng, Lu Xu, Runsheng Yu, Ruolan Cao, Yiming Du, Tianyu Han, Yuan Hu, Zhen Li, Shuai Liu, Simeng Ma, Shuo Pan, Junru Xiao, Ning Xu, Ming Yang, Rong Ye, Yikun Yu, Jinhua Zhang, Ruiheng Zhang, Wenqi Zhang, Wenqiao Zhu, Longsheng Zou, Lin Lu, Yingyi Wang, and Yiming Wu. Seed liveinterpret 2.0: End-to-end simultaneous speech-to-speech translation with your voice. *arXiv preprint arXiv:2507.17527*, 2025. URL https://arxiv.org/abs/2507.17527.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proc. Interspeech 2020*, 2020.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in rl for large language models. *arXiv preprint arXiv:2505.23564*, 2025. URL https://arxiv.org/abs/2505.23564. NeurIPS 2025.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. *arXiv preprint arXiv:2410.01679*, 2024. URL https://arxiv.org/abs/2410.01679.

Tom Labiausse, Laurent Mazaré, Edouard Grave, Alexandre Défossez, and Neil Zeghidour. High-fidelity simultaneous speech-to-speech translation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32116–32129, 2025. URL https://proceedings.mlr.press/v267/labiausse25a.html.

Tom Labiausse, Romain Fabre, Yannick Estève, Alexandre Défossez, and Neil Zeghidour. Simultaneous speech-to-speech translation without aligned data. *arXiv preprint arXiv:2602.11072*, 2026. URL https://arxiv.org/abs/2602.11072.

Siqi Ouyang, Shuoyang Ding, Oleksii Hrinchuk, Vitaly Lavrukhin, Brian Yan, Boris Ginsburg, and Lei Li. Hierarchical policy optimization for simultaneous translation of unbounded speech. *arXiv preprint arXiv:2604.21045*, 2026. URL https://arxiv.org/abs/2604.21045.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. Over-generation cannot be rewarded: Length-adaptive average lagging for simultaneous speech translation. In *Proceedings of the Third Workshop on Automatic Simultaneous Translation*, 2022. URL https://aclanthology.org/2022.autosimtrans-1.2/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, Teven Le Scao, Yihan Wang, Adam Yang, Alexander H. Liu, et al. Magistral. *arXiv preprint arXiv:2506.10910*, 2025. URL https://arxiv.org/abs/2506.10910. Mistral AI.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Proc. Interspeech 2022*, 2022.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Seungwoo Shin, Dongkyu Ahn, Junhyeok Kim, and Seungjae Jeon. No verifiable reward for prosody: Toward preference-guided prosody learning in tts. *arXiv preprint arXiv:2509.18531*, 2025. URL https://arxiv.org/abs/2509.18531.

Haoxu Wang, Biao Tian, Weiqing Li, Xiang Lv, Han Zhao, and Xiangang Li. Flowtts-grpo: Online reinforcement learning with multi-objective reward optimization for flow-matching based text-to-speech. *arXiv preprint arXiv:2606.23190*, 2026. URL https://arxiv.org/abs/2606.23190.

Tianning Xu, Zhen Huang, Jian Sun, Shanbo Cheng, and Wai Lam. Seqpo-simt: Sequential policy optimization for simultaneous machine translation. *arXiv preprint arXiv:2505.20622*, 2025. URL https://arxiv.org/abs/2505.20622. Findings of ACL 2025.

Donglei Yu, Yang Xie, Yanzhi Chen, and Yang Feng. Simulpl: Aligning human preferences in simultaneous machine translation. *arXiv preprint arXiv:2502.00634*, 2025. URL https://arxiv.org/abs/2502.00634. ICLR 2025.

Guohui Zhang, XiaoXiao Ma, Jie Huang, Hang Xu, Hu Yu, Siming Fu, Yuming Li, Zeyue Xue, Lin Song, Haoyang Huang, Nan Duan, and Feng Zhao. Omninft: Modality-wise omni diffusion reinforcement for joint audio-video generation. *arXiv preprint arXiv:2605.12480*, 2026. URL https://arxiv.org/abs/2605.12480.

Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. Streamspeech: Simultaneous speech-to-speech translation with multi-task learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL https://aclanthology.org/2024.acl-long.485/.

A HYPERPARAMETERS

Table 4 reports hyperparameters shared across all RL variants; Table 5 reports settings that differ by variant. LoRA adapters (rank 16, 4-bit NF4 quantization) target 36 projection modules spanning the temporal- and depth-attention blocks, the depformer input projection, and the output projection heads, parameterizing 11,124,736 trainable weights (0.40% of the model backbone), identical across every variant.

Table 4: Hyperparameters shared across all RL variants.

Hyperparameter	Value
Learning rate	1×10^{-5}
PPO/GRPO clip range	0.2
Sampling temperature / top-k	0.8 / 250
Stream weighting (c_0 / c_q)	100 / 1
Process-reward mixing coefficient (α)	0.4
Process-reward word positions (n_w)	8
LoRA rank / quantization	16 / 4-bit NF4

Table 5: Variant-specific hyperparameters and training length.

Variant	Group size	KL coeff.	Value head	Steps	Reported ckpt.
GRPO	4	—	—	1000	step 300
KL-Regularized GRPO	4	0.02	—	600	step 600
PPO+GAE	1	0.02	2048→512→1	600	step 600
Matched-Reward PPO+GAE	1	0.02	2048→512→1	600	step 600
PPO+GAE (group size = 4)	4	0.02	2048→512→1	600	step 600
Magnitude-Gated GRPO	4	—	2048→1	600	step 600
Additive Value-Baseline GRPO	4	—	2048→1	600	step 600
Audio-Native Value Target	4	—	2048→1	600	step 600
Segment-Level Critic-Free	4	—	—	600	step 600
GRPO (group size = 2)	2	—	—	300	step 300

Table 6: Baseline, GRPO, and PPO+GAE pairwise comparisons. $\alpha = 0.00417$ (3 comparisons $\times$ 4 metrics).

Metric	Comparison	Mean diff. [95% CI]	Significant
ASR-BLEU	Baseline vs GRPO	+1.146 [−0.78, +3.05]	No
ASR-BLEU	Baseline vs PPO+GAE	−1.079 [−3.15, +0.95]	No
ASR-BLEU	GRPO vs PPO+GAE	−2.225 [−4.33, −0.20]	No
UTMOS	Baseline vs GRPO	+0.038 [−0.03, +0.11]	No
UTMOS	Baseline vs PPO+GAE	+0.0005 [−0.07, +0.07]	No
UTMOS	GRPO vs PPO+GAE	−0.038 [−0.11, +0.03]	No
Speaker sim.	Baseline vs GRPO	+0.0254 [+0.008, +0.042]	Yes
Speaker sim.	Baseline vs PPO+GAE	+0.0019 [−0.016, +0.020]	No
Speaker sim.	GRPO vs PPO+GAE	−0.0235 [−0.040, −0.007]	No
LAAL (s)	Baseline vs GRPO	−0.305 [−0.474, −0.138]	Yes
LAAL (s)	Baseline vs PPO+GAE	+0.057 [−0.082, +0.197]	No
LAAL (s)	GRPO vs PPO+GAE	+0.362 [+0.226, +0.504]	Yes

Table 7: Magnitude-Gated GRPO vs. baseline and GRPO@600. $\alpha = 0.00625$ (2 comparisons $\times$ 4 metrics).

Metric	Comparison	Mean diff. [95% CI]	Significant
ASR-BLEU	vs Baseline	−1.539 [−3.61, +0.48]	No
ASR-BLEU	vs GRPO@600	+0.775 [−1.11, +2.67]	No
UTMOS	vs Baseline	−0.378 [−0.465, −0.293]	Yes
UTMOS	vs GRPO@600	−0.008 [−0.092, +0.075]	No
Speaker sim.	vs Baseline	−0.086 [−0.104, −0.068]	Yes
Speaker sim.	vs GRPO@600	+0.007 [−0.010, +0.025]	No
LAAL (s)	vs Baseline	+0.648 [+0.472, +0.831]	Yes
LAAL (s)	vs GRPO@600	−0.014 [−0.101, +0.072]	No

Table 8: KL-Regularized GRPO vs. baseline, GRPO@600, and PPO+GAE@600. $\alpha = 0.00417$ (3 comparisons $\times$ 4 metrics).

Metric	Comparison	Mean diff. [95% CI]	Significant
ASR-BLEU	vs Baseline	−1.508 [−3.63, +0.60]	No
ASR-BLEU	vs GRPO@600	+0.805 [−1.03, +2.58]	No
ASR-BLEU	vs PPO+GAE@600	−2.587 [−4.86, −0.32]	No
UTMOS	vs Baseline	−0.3865 [−0.470, −0.306]	Yes
UTMOS	vs GRPO@600	−0.0161 [−0.089, +0.058]	No
UTMOS	vs PPO+GAE@600	−0.3860 [−0.466, −0.305]	Yes
Speaker sim.	vs Baseline	−0.0785 [−0.097, −0.060]	Yes
Speaker sim.	vs GRPO@600	+0.0146 [−0.004, +0.033]	No
Speaker sim.	vs PPO+GAE@600	−0.0767 [−0.095, −0.058]	Yes
LAAL (s)	vs Baseline	+0.6702 [+0.490, +0.847]	Yes
LAAL (s)	vs GRPO@600	+0.0079 [−0.094, +0.110]	No
LAAL (s)	vs PPO+GAE@600	+0.7268 [+0.555, +0.904]	Yes

Table 9: Matched-Reward PPO+GAE vs. baseline, GRPO@600, and original PPO+GAE@600. $\alpha = 0.00417$ (3 comparisons $\times$ 4 metrics).

Metric	Comparison	Mean diff. [95% CI]	Significant
UTMOS	vs Baseline	−0.0098 [−0.082, +0.059]	No
UTMOS	vs GRPO@600	+0.3606 [+0.282, +0.442]	Yes
UTMOS	vs PPO+GAE@600	−0.0094 [−0.067, +0.046]	No
Speaker sim.	vs Baseline	−0.0023 [−0.021, +0.017]	No
Speaker sim.	vs GRPO@600	+0.0909 [+0.071, +0.111]	Yes
Speaker sim.	vs PPO+GAE@600	−0.0004 [−0.016, +0.015]	No
LAAL (s)	vs Baseline	−0.0685 [−0.205, +0.063]	No
LAAL (s)	vs GRPO@600	−0.7308 [−0.887, −0.581]	Yes
LAAL (s)	vs PPO+GAE@600	−0.0119 [−0.109, +0.087]	No

B Full Significance Testing Results

All tests use paired bootstrap resampling (10,000 resamples) and the Wilcoxon signed-rank test, Bonferroni-corrected within each family of comparisons; a result is significant only where both tests agree after correction.

Table 10: Segment-Level Critic-Free vs. baseline, GRPO@300, and PPO+GAE@600. $\alpha = 0.00417$ (3 comparisons $\times$ 4 metrics).

Metric	Comparison	Mean diff. [95% CI]	Significant
ASR-BLEU	vs Baseline	+0.99 [−0.87, +2.90]	No
ASR-BLEU	vs GRPO@300	+2.14 [+0.25, +4.10]	No
ASR-BLEU	vs PPO+GAE@600	−0.09 [−2.09, +1.97]	No
UTMOS	vs Baseline	−0.308 [−0.391, −0.225]	Yes
UTMOS	vs GRPO@300	−0.270 [−0.353, −0.188]	Yes
UTMOS	vs PPO+GAE@600	−0.307 [−0.386, −0.230]	Yes
Speaker sim.	vs Baseline	−0.0818 [−0.0997, −0.0638]	Yes
Speaker sim.	vs GRPO@300	−0.0564 [−0.0740, −0.0384]	Yes
Speaker sim.	vs PPO+GAE@600	−0.0800 [−0.0984, −0.0614]	Yes
LAAL (s)	vs Baseline	+0.6971 [+0.5052, +0.8972]	Yes
LAAL (s)	vs GRPO@300	+0.3920 [+0.2342, +0.5548]	Yes
LAAL (s)	vs PPO+GAE@600	+0.7537 [+0.5644, +0.9525]	Yes

Table 11: Segment-Level Critic-Free vs. Additive Value-Baseline GRPO and Audio-Native Value Target, both @600. Reported for the qualitative claims made in the main text; evaluated at the same $\alpha = 0.00417$ threshold.

Metric	Comparison	Mean diff. [95% CI]	Significant
ASR-BLEU	vs Additive Value-Baseline	+2.18 [+0.15, +4.27]	No
ASR-BLEU	vs Audio-Native Value Target	+3.44 [+1.56, +5.40]	Yes
UTMOS	vs Additive Value-Baseline	+0.003 [−0.085, +0.091]	No
UTMOS	vs Audio-Native Value Target	+0.191 [+0.099, +0.284]	Yes
Speaker sim.	vs Additive Value-Baseline	−0.0205 [−0.0381, −0.0031]	No
Speaker sim.	vs Audio-Native Value Target	+0.0368 [+0.0196, +0.0545]	Yes
LAAL (s)	vs Additive Value-Baseline	+0.0782 [−0.0627, +0.2259]	No
LAAL (s)	vs Audio-Native Value Target	+0.0022 [−0.1486, +0.1573]	No

Table 12: PPO+GAE at group size 4 vs. baseline, original PPO+GAE (group size 1), and GRPO@600. $\alpha = 0.00417$ (3 comparisons $\times$ 4 metrics).

Metric	Comparison	Mean diff. [95% CI]	Significant
ASR-BLEU	vs Baseline	−0.12 [−1.99, +1.71]	No
ASR-BLEU	vs PPO+GAE (gs=1)	−1.20 [−3.04, +0.57]	No
ASR-BLEU	vs GRPO@600	+2.19 [+0.60, +3.86]	No
UTMOS	vs Baseline	−0.002 [−0.065, +0.061]	No
UTMOS	vs PPO+GAE (gs=1)	−0.002 [−0.060, +0.055]	No
UTMOS	vs GRPO@600	+0.368 [+0.289, +0.446]	Yes
Speaker sim.	vs Baseline	−0.0002 [−0.0166, +0.0158]	No
Speaker sim.	vs PPO+GAE (gs=1)	+0.0017 [−0.0162, +0.0186]	No
Speaker sim.	vs GRPO@600	+0.0930 [+0.0766, +0.1096]	Yes
LAAL (s)	vs Baseline	−0.0866 [−0.2223, +0.0487]	No
LAAL (s)	vs PPO+GAE (gs=1)	−0.0300 [−0.1162, +0.0555]	No
LAAL (s)	vs GRPO@600	−0.7489 [−0.9113, −0.5948]	Yes

Table 13: Each c_q value vs. baseline (no RL) at step 600. $\alpha = 0.003125$ (4 comparisons $\times$ 4 metrics).

Metric	Comparison	Mean diff. [95% CI]	Significant
ASR-BLEU	c_q=0 vs Baseline	−2.61 [−4.58, −0.64]	No
ASR-BLEU	c_q=0.1 vs Baseline	−1.11 [−3.33, +1.20]	No
ASR-BLEU	c_q=0.5 vs Baseline	−1.28 [−3.12, +0.62]	No
ASR-BLEU	c_q=1 vs Baseline	−2.31 [−4.15, −0.41]	No
UTMOS	c_q=0 vs Baseline	−0.661 [−0.767, −0.556]	Yes
UTMOS	c_q=0.1 vs Baseline	−0.618 [−0.722, −0.512]	Yes
UTMOS	c_q=0.5 vs Baseline	−0.326 [−0.407, −0.244]	Yes
UTMOS	c_q=1 vs Baseline	−0.370 [−0.450, −0.292]	Yes
Speaker sim.	c_q=0 vs Baseline	−0.1070 [−0.1272, −0.0866]	Yes
Speaker sim.	c_q=0.1 vs Baseline	−0.1067 [−0.1252, −0.0885]	Yes
Speaker sim.	c_q=0.5 vs Baseline	−0.0715 [−0.0889, −0.0543]	Yes
Speaker sim.	c_q=1 vs Baseline	−0.0932 [−0.1112, −0.0749]	Yes
LAAL (s)	c_q=0 vs Baseline	+0.5583 [+0.3765, +0.7408]	Yes
LAAL (s)	c_q=0.1 vs Baseline	+0.5423 [+0.3728, +0.7157]	Yes
LAAL (s)	c_q=0.5 vs Baseline	+0.5833 [+0.4051, +0.7595]	Yes
LAAL (s)	c_q=1 vs Baseline	+0.6623 [+0.4888, +0.8470]	Yes

Table 14: c_q dose-response pairwise comparisons at step 600. $\alpha = 0.00208$ (6 comparisons $\times$ 4 metrics).

Metric	Comparison	Mean diff. [95% CI]	Significant
ASR-BLEU	c_q=0 vs c_q=0.1	-1.51 [-3.69, $+0.62$]	No
ASR-BLEU	c_q=0 vs c_q=0.5	-1.34 [-3.14, $+0.39$]	No
ASR-BLEU	c_q=0 vs c_q=1	-0.30 [-2.00, $+1.37$]	No
ASR-BLEU	c_q=0.1 vs c_q=0.5	+0.17 [-1.75, $+2.13$]	No
ASR-BLEU	c_q=0.1 vs c_q=1	+1.21 [-0.88, $+3.36$]	No
ASR-BLEU	c_q=0.5 vs c_q=1	+1.04 [-0.77, $+2.88$]	No
UTMOS	c_q=0 vs c_q=0.1	-0.043 [-0.124, $+0.038$]	No
UTMOS	c_q=0 vs c_q=0.5	-0.335 [-0.418, -0.252]	Yes
UTMOS	c_q=0 vs c_q=1	-0.290 [-0.375, -0.206]	Yes
UTMOS	c_q=0.1 vs c_q=0.5	-0.292 [-0.376, -0.209]	Yes
UTMOS	c_q=0.1 vs c_q=1	-0.248 [-0.332, -0.167]	Yes
UTMOS	c_q=0.5 vs c_q=1	+0.044 [-0.032, $+0.121$]	No
Speaker sim.	c_q=0 vs c_q=0.1	-0.0003 [-0.0160, $+0.0156$]	No
Speaker sim.	c_q=0 vs c_q=0.5	-0.0355 [-0.0520, -0.0189]	Yes
Speaker sim.	c_q=0 vs c_q=1	-0.0138 [-0.0306, $+0.0029$]	No
Speaker sim.	c_q=0.1 vs c_q=0.5	-0.0352 [-0.0505, -0.0193]	Yes
Speaker sim.	c_q=0.1 vs c_q=1	-0.0135 [-0.0313, $+0.0038$]	No
Speaker sim.	c_q=0.5 vs c_q=1	+0.0217 [$+0.0042$, $+0.0396$]	No
LAAL (s)	c_q=0 vs c_q=0.1	+0.0160 [-0.0857, $+0.1214$]	No
LAAL (s)	c_q=0 vs c_q=0.5	-0.0250 [-0.1343, $+0.0862$]	No
LAAL (s)	c_q=0 vs c_q=1	-0.1040 [-0.2039, -0.0064]	No
LAAL (s)	c_q=0.1 vs c_q=0.5	-0.0410 [-0.1375, $+0.0567$]	No
LAAL (s)	c_q=0.1 vs c_q=1	-0.1200 [-0.2165, -0.0242]	No
LAAL (s)	c_q=0.5 vs c_q=1	-0.0790 [-0.1857, $+0.0277$]	No

C REPRODUCIBILITY

All experiments were conducted on a single NVIDIA RTX 3090 GPU (24GB). Training and evaluation data are drawn from Europarl-ST (French-English) and Google FLEURS (French, Spanish-English), both publicly available. Code and configuration files sufficient to reproduce all reported experiments will be released upon publication.