SELF-REWARDING GRPO FOR INSTRUCTION-GUIDED ZERO-SHOT SPEECH SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Following natural-language instructions while preserving an unseen speaker's identity remains challenging for zero-shot text-to-speech. We present ACE-Instruct-TTS and Self-Rewarding Group Relative Policy Optimization (SR-GRPO), a post-training method that derives instruction rewards from the TTS model itself rather than an external audio understanding model. SR-GRPO uses a frozen supervised fine-tuning (SFT) checkpoint to score the same generated speech tokens with and without the instruction, holding the target text and speaker conditioning fixed. Inspired by conditional pointwise mutual information, the length-normalized log-likelihood ratio provides a continuous measure of relative compatibility with the instruction. We combine this signal with speaker-similarity and duration rewards. We also introduce ACE-Instruct-TTS-Eval-ZH, a 650-sample Chinese benchmark evaluating instruction following, speaker preservation, and intelligibility across six categories, including compositional and conflicting conditions. SR-GRPO improves instruction-following rate from 67.08% to 75.23%, the highest among evaluated systems, and increases mean human-rated instruction-following scores from 3.64 to 3.90. Reward ablations favor the proposed likelihood contrast over conditional-only, rank-based, and external binary rewards. Speaker similarity remains above all evaluated external baselines but decreases relative to SFT, while character error rate remains comparable. Beyond this benchmark, SR-GRPO improves the Chinese InstructTTSEval macro-average score from 79.2% to 81.2% and improves joint success across three SpeechEditBench tasks under an adapted re-synthesis protocol. On a second backbone, CosyVoice3, SR-GRPO improves instruction-following rate by 5.54 percentage points, with a similar speaker-preservation trade-off. We train instruction following only in Mandarin and leave multilingual instruction following to future work.

1 INTRODUCTION

Zero-shot text-to-speech (TTS) has made substantial progress in high-fidelity voice cloning Wang et al. (2023); Anastassiou et al. (2024); Du et al. (2025). Practical applications also require control over delivery, including emotion, speaking style, and prosody. Early controllable TTS systems use categorical labels or learned style tokens Wang et al. (2018); Cai et al. (2021), whereas recent systems support natural-language instructions Hu et al. (2026a); Huang et al. (2026); Hu et al. (2026b). Combining instruction control with zero-shot voice cloning requires modifying delivery attributes while preserving an unseen speaker identity. Recent systems support both reference-speech and instruction conditioning Zhou et al. (2024); Du et al. (2025); Zhou et al. (2026), but reliable control remains difficult when instructions combine multiple attributes or conflict with cues in the target text or reference speech.

Reinforcement learning (RL) can improve instruction adherence beyond supervised fine-tuning (SFT). Recent work explores preference optimization and Group Relative Policy Optimization (GRPO) for controllable speech generation Rafailov et al. (2024); Shao et al. (2024); Chen et al. (2026a), but their effectiveness depends critically on the reward. Some approaches use external audio understanding models to assess instruction adherence Chen et al. (2026a), requiring an additional evaluator for sampled rollouts. Binary feedback cannot distinguish degrees of partial compliance; when instruction rewards are identical within a group, that component provides no group-relative

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.

learning signal. This motivates a central question: *Can a TTS model itself provide a graded reward for instruction adherence without an external instruction evaluator?*

We present **ACE-Instruct-TTS**, a decoder-only speech language model for instruction-based voice design and zero-shot voice cloning, and **Self-Rewarding GRPO (SR-GRPO)**. Using a frozen SFT scorer π_{ref} derived from the model itself rather than an external judge, SR-GRPO compares each rollout's channel-weighted mean negative log-likelihood (NLL) with and without the instruction: $r_{\text{inst}} = \mathcal{L}_{\text{no-inst}} - \mathcal{L}_{\text{inst}}$, while keeping the target text, speaker conditioning, and scored speech tokens fixed. Inspired by conditional pointwise mutual information (CPMI), this contrast measures relative instruction compatibility, using instruction-free prediction as a baseline rather than an explicit instruction marginal. The frozen scorer provides stationary, continuous feedback that retains score differences beyond binary labels or ranks. We combine it with speaker-similarity and duration rewards through GDPO's reward-decoupled group normalization Liu et al. (2026).

Existing voice-design benchmarks such as InstructTTSEval and MINT-Bench do not directly assess instruction following with reference-speaker preservation Huang et al. (2025); Chen et al. (2026b). We introduce **ACE-Instruct-TTS-Eval-ZH**, a 650-sample Chinese benchmark pairing reference speech, target texts, and instructions to evaluate instruction following, speaker similarity, and intelligibility. Its six categories cover emotion and tone, speaking style, dialects and accents, acoustic controls, compositional conditions, and conflicts with the target text or reference delivery.

On ACE-Instruct-TTS-Eval-ZH, SR-GRPO raises instruction-following rate (IFR) from 67.08% to 75.23%, leading all evaluated systems, and mean instruction-following MOS from 3.64 to 3.90. Speaker similarity decreases from SFT but exceeds all external baselines, with comparable character error rate (1.97% versus 2.01%). Conditional-only, rank-based, and external binary rewards achieve lower IFRs of 68.77%, 69.38%, and 67.69%, respectively. SR-GRPO also improves the Chinese InstructTTSEval macro average from 79.2% to 81.2% and joint success on SpeechEditBench's Emotion, Style, and Prosody tasks under our adapted reference-conditioned re-synthesis protocol. Beyond ACE-Instruct-TTS, applying SR-GRPO to CosyVoice3 improves IFR from 66.31% to 71.85%, with a similar adherence–speaker-preservation trade-off. These results provide evidence of benefits across evaluation settings and on a second TTS backbone. Benchmark materials, evaluation code, and our SR-GRPO implementation for the public CosyVoice3 backbone will be released subject to applicable redistribution permissions. Our main contributions are:

- **CPMI-inspired self-rewarding.** We introduce SR-GRPO, deriving a continuous instruction reward from instruction-conditioned versus instruction-free likelihoods under a frozen SFT model, without an external audio understanding model for instruction scoring.
- **Instruction-controlled zero-shot TTS.** We develop ACE-Instruct-TTS and demonstrate SFT-to-RL improvements across three evaluation settings. Candidate-scoring analyses and reward ablations support the likelihood contrast and characterize the trade-off between instruction adherence and speaker preservation. Audio samples are available on our anonymous demo page: `https://anonymous.4open.science/w/ace-instruct-tts/`.
- **A benchmark for instruction-guided voice cloning.** We construct ACE-Instruct-TTS-Eval-ZH to jointly evaluate instruction following, speaker preservation, and intelligibility across six categories, including compositional and conflicting conditions.

2 Related Work

Instruction-guided speech synthesis. Controllable TTS has evolved from emotion labels and style tokens Wang et al. (2018); Cai et al. (2021) to natural-language conditioning for voice design and expressive generation Hu et al. (2026a); Huang et al. (2026); Hu et al. (2026b). VoxInstruct, CosyVoice 3, and VoxCPM2 combine instructions with reference-speech conditioning to control delivery while preserving speaker identity Zhou et al. (2024); Du et al. (2025); Zhou et al. (2026). ACE-Instruct-TTS follows this setting, focusing on post-training rewards for instruction adherence and speaker preservation.

GRPO in instruction-guided TTS. GRPO optimizes sampled outputs using group-relative reward advantages, and has been extended to instruction-guided speech synthesis to improve control-

lability Shao et al. (2024). FlexiVoice applies GRPO post-training to controllable speech with an external audio understanding model for instruction evaluation Chen et al. (2026a). *Listen, Critique, and Refine* explores self-refinement RL for instruction-following speech yet still relies on a separate critique module rather than deriving rewards from the TTS model itself Yu et al. (2026). Other GRPO-based TTS works mainly optimize intelligibility and speaker similarity Sun et al. (2025); Liu et al. (2025). SR-GRPO instead derives continuous instruction rewards from a frozen SFT reference, without an external instruction evaluator, and retains auxiliary speaker-similarity and duration rewards.

Self-rewarding and mutual-information-based scoring. Self-rewarding methods use generative models to provide alignment feedback, such as preference judgments Yuan et al. (2025). SR-GRPO instead derives rewards from frozen SFT token likelihoods, drawing on mutual-information-based generation and PMI scoring to measure input-specific compatibility via likelihood contrasts Li et al. (2016); Holtzman et al. (2022). With text and speaker context fixed, we contrast rollout log-likelihoods with and without instructions. The length-normalized contrast yields a CPMI-inspired reward, though the instruction-free baseline is not an explicit instruction marginal. Relatedly, classifier-free guidance (CFG) combines conditional and unconditional predictions at inference Ho & Salimans (2022), including next-token guidance in autoregressive language and audio models Sanchez et al. (2023); Copet et al. (2024). SR-GRPO instead uses the contrast as a sequence-level reward to update the policy rather than the decoding distribution.

3 Overview of ACE-Instruct-TTS

3.1 ACE-Instruct-TTS Architecture

Backbone. ACE-Instruct-TTS builds on ACE-TTS-Base, an in-house zero-shot TTS model that combines a decoder-only LLM with discrete residual vector quantization (RVQ) tokens. The backbone is initialized from Qwen3-1.7B-Instruct and models $C = 8$ RVQ channels using delay-pattern interleaved prediction Copet et al. (2024). Speech is discretized using the DualCodec dual-stream tokenizer Li et al. (2025): the first RVQ layer primarily encodes semantic content, while the remaining layers provide acoustic detail. A flow-matching model Lipman et al. (2023) followed by BigVGAN-v2 Lee et al. (2023) reconstructs the waveform. Speaker identity is conditioned on an embedding extracted from the reference audio using WavLM-large fine-tuned for speaker verification Chen et al. (2022). To enable instruction control, we insert the instruction text between the target text and the speaker embedding placeholder. An architecture diagram is provided in Appendix A.2 (Figure 3).

Training. Training proceeds in three stages. *Stage 1: joint voice design and zero-shot voice cloning.* We jointly fine-tune ACE-TTS-Base on in-house instruction-only voice-design data (without reference speaker audio) and instruction-free voice-cloning data. *Stage 2: instruction-controlled voice cloning.* From the Stage 1 checkpoint, we jointly train voice cloning with reference speaker audio, with and without instructions. Instruction-controlled training first uses a large corpus with noisy instruction–audio alignment, then a high-quality corpus with more reliable alignment to refine instruction adherence, retaining instruction-free data throughout. The resulting model is denoted as ACE-Instruct-TTS-SFT. *Stage 3: self-rewarding policy optimization.* We further optimize ACE-Instruct-TTS-SFT using SR-GRPO (Section 3.2).

3.2 CPMI-Inspired Self-Rewarding

Formulation of Conditional PMI. Let $z = \{x_{\text{text}}, c_{\text{spk}}\}$ denote the context shared by all rollouts for a given prompt, consisting of the target text and speaker conditioning. Let c_{inst} denote the instruction and y a generated speech-token sequence. Conditional pointwise mutual information (CPMI) is defined as

$$\mathrm{PMI}(c_{\text{inst}}; y \mid z) = \log p(y \mid c_{\text{inst}}, z) - \log p(y \mid z), \tag{1}$$

where $p(y \mid z)$ denotes the instruction-marginal distribution. CPMI measures the change in the log-likelihood of a rollout when the instruction is specified, relative to this marginal distribution, while holding the target text and speaker conditioning fixed.

Self-Rewarding via a Frozen Reference Model. We use a frozen copy of ACE-Instruct-TTS-SFT, π_{ref}, as the scorer. Both supervised training stages include instruction-free voice cloning, providing training support for scoring speech without instruction input. For each rollout y, we perform two teacher-forced passes: one with the instruction and one without it, while retaining the same target text, speaker conditioning, and speech-token sequence. Here, $\pi_{\text{ref}}(y \mid z)$ denotes the model prediction with the instruction omitted. It serves as a comparison baseline, rather than an explicitly marginalized distribution over instructions. Freezing the scorer keeps the reward function stationary throughout policy optimization. Only valid speech tokens contribute to scoring; conditioning and padding tokens are excluded. We compute a weighted average of the mean NLLs over the eight RVQ channels:

$$\mathcal{L}_{\text{inst}} = -\sum_{j=1}^{8} \frac{\alpha_j}{T_j} \sum_{t \in \mathcal{V}_j} \log \pi_{\text{ref}}(y_{t,j} \mid h_{t,j}, c_{\text{inst}}, z),$$

$$\mathcal{L}_{\text{no-inst}} = -\sum_{j=1}^{8} \frac{\alpha_j}{T_j} \sum_{t \in \mathcal{V}_j} \log \pi_{\text{ref}}(y_{t,j} \mid h_{t,j}, z), \tag{2}$$

where $\mathcal{V}_j$ contains the valid positions in channel j, $T_j = |\mathcal{V}_j|$, and $h_{t,j}$ denotes the preceding speech-token context under the delay pattern. The normalized channel weights are $\boldsymbol{\alpha} = (12, 8, 6, 4, 2, 2, 1, 1)/36$, matching the SFT loss weights. Both passes score identical speech tokens with the same channel weights and valid-token masks. The instruction reward is their difference:

$$r_{\text{inst}} = \mathcal{L}_{\text{no-inst}} - \mathcal{L}_{\text{inst}} = \sum_{j=1}^{8} \frac{\alpha_j}{T_j} \sum_{t \in \mathcal{V}_j} \log \frac{\pi_{\text{ref}}(y_{t,j} \mid h_{t,j}, c_{\text{inst}}, z)}{\pi_{\text{ref}}(y_{t,j} \mid h_{t,j}, z)}. \tag{3}$$

The reward measures how much the instruction reduces the channel-weighted mean NLL relative to the instruction-free baseline, capturing relative instruction compatibility rather than absolute likelihood. Normalization by T_j places each channel's contrast on a per-token scale.

3.3 Auxiliary Rewards and Training objective

Auxiliary rewards. We use duration and speaker-similarity rewards. For generated and paired target durations L_{gen} and $L_{\text{tgt}} > 0$, the clipped duration reward is $r_{\text{len}} = \max(-1, 1 - 2|\rho|)$, where $\rho = (L_{\text{gen}} - L_{\text{tgt}})/L_{\text{tgt}}$. The timbre reward is $r_{\text{timbre}} = \cos(E(D(y)), E(c_{\text{spk}}))$, where $D(y)$ is the decoded waveform, c_{spk} is the reference audio, and E is a WavLM-large-based speaker embedding extractor Chen et al. (2022).

Training objective. For each prompt (c_{inst}, z), we sample $G = 8$ rollouts $\{y_i\}_{i=1}^{G}$ from π_θ and, following GDPO Liu et al. (2026), standardize each reward head $k \in \{\text{inst}, \text{timbre}, \text{len}\}$ independently within the group before fusion: $A_i = \sum_k \lambda_k \, (r_i^{(k)} - \mu^{(k)})/(\sigma^{(k)} + \epsilon)$, where $\mu^{(k)}$ and $\sigma^{(k)}$ are group statistics and $(\lambda_{\text{inst}}, \lambda_{\text{timbre}}, \lambda_{\text{len}}) = (0.6, 0.3, 0.1)$. we use $\epsilon = 10^{-6}$. A_i is broadcast to every generated token–channel cell (s, j) of the delay-pattern RVQ output Copet et al. (2024), and the policy-gradient and KL terms share the same mask and channel weights:

$$\mathcal{L}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \frac{1}{Z_i} \sum_{s,j} w_j \, m_{i,s,j} \Big[- A_i \log \pi_\theta(y_{i,s,j} \mid h_{i,s,j}, c_{\text{inst}}, z) + \beta_t \, \hat{k}_{i,s,j}^{(3)} \Big], \tag{4}$$

where $m_{i,s,j}$ masks generated cells, $w_j \propto (12, 8, 6, 4, 2, 2, 1, 1)$, $Z_i = \sum_{s,j} w_j m_{i,s,j}$ normalizes for length, and $\hat{k}_{i,s,j}^{(3)}$ is the k_3 KL estimator against the frozen π_{ref} with adaptive coefficient β_t. Because each rollout batch receives a single gradient step, the importance ratio is identically one, so the clipped GRPO surrogate reduces to this REINFORCE form with a group-standardized baseline (details in Appendix A.3).

4 ACE-Instruct-TTS-Eval-ZH

Evaluating instruction-controlled zero-shot voice cloning requires jointly assessing instruction adherence and reference-speaker preservation. We introduce **ACE-Instruct-TTS-Eval-ZH**, a 650-

sample Chinese benchmark covering expressive delivery, regional speech variation, acoustic controls, and compositional or conflicting conditions. Each sample pairs a reference clip and its transcript from *seed-tts-eval* Anastassiou et al. (2024) with a natural-language instruction and target text, generated by gpt-5.6-luna and human-verified (Appendix A.1). Table 1 lists the six categories and their counts. Emotion and speaking style follow hierarchical taxonomies; dialect and accent instructions specify regional characteristics; acoustic controls cover speaking rate, pitch, volume, and whispering. Compositional conditions require jointly realizing two to four attributes, while conflicting conditions pair instructions with incongruent target-text content or reference delivery, testing whether models prioritize the requested delivery over competing contextual cues while preserving speaker identity. The benchmark jointly evaluates instruction following, speaker similarity, and intelligibility, exposing trade-offs that instruction-only evaluation may overlook. Metric definitions and evaluation protocols are provided in Section 5.1.

Table 1: Composition of ACE-Instruct-TTS-Eval-ZH. Details are provided in Appendix A.1.

Category	**Count**	**Coverage**
Emotion & Tone	180	Seven emotion groups with fine-grained subtypes.
Speaking Style	100	Four register groups comprising eight substyles.
Dialects & Accents	64	Regional varieties, e.g., Cantonese, Shanghainese, Sichuan, and Northeastern speech.
Acoustic Controls	50	Speaking rate, pitch, volume, and whispering.
Compositional Conditions	156	Joint control of two to four attributes.
Conflicting Conditions	100	Instructions conflicting with target-text content or reference delivery.
Total	**650**	

5 Experiments

5.1 Setup

Data. Stage 1 uses 20,000 h of in-house Mandarin instruction-only data and 4,000 h of instruction-free voice-cloning data, equally split between Mandarin and English. Stage 2 first combines 10,000 h of Mandarin instruction-controlled voice-cloning data with noisy instruction–audio alignment and the same 4,000 h instruction-free corpus. It then uses 3,000 h of verified Mandarin data with more reliable alignment, mixed with a 1,000 h subset of the instruction-free corpus. Stage 3 uses 10,000 instruction–utterance pairs sampled exclusively from the verified Stage 2 data. ACE-Instruct-TTS-Eval-ZH speakers, utterances, and instructions are held out from all training stages.

Benchmarks. On **ACE-Instruct-TTS-Eval-ZH** (Section 4), we report IFR, IFMOS, SIM, and CER. On the 1,000-sample Chinese subset of **InstructTTSEval** Huang et al. (2025), we evaluate voice design without speaker references, reporting Acoustic-Parameter Specification (APS), Descriptive-Style Directive (DSD), Role-Play (RP), and their macro average. For **SpeechEdit-Bench** Zhang et al. (2026), we evaluate Emotion, Style, and Prosody without editing-specific training. We report target success and joint success. Since our instruction-training data are entirely in Chinese, we translate English instructions into Chinese while retaining the original transcript and output languages. Published baselines use untranslated instructions, limiting direct cross-system comparisons.

Baselines. On ACE-Instruct-TTS-Eval-ZH, we compare against two open-source systems, CosyVoice3 Du et al. (2025) and VoxCPM2 Zhou et al. (2026), and two commercial systems, Qwen-Audio-3.0-TTS-Plus Xiang et al. (2026) and MiMo-V2.5-TTS Xiaomi (2026), accessed via their public APIs on August 19, 2026. We exclude FlexiVoice Chen et al. (2026a), which has no public checkpoint. For InstructTTSEval and SpeechEditBench, we compare against the task-specific systems listed in Tables 3 and 4, respectively. For the external benchmarks, we use published baseline results.

Metrics. **IFR** on ACE-Instruct-TTS-Eval-ZH follows the InstructTTSEval evaluation procedure Huang et al. (2025), using Gemini 2.5 Pro with majority voting over three binary judgments per

generated utterance. **SIM** (speaker similarity) and **CER** (character error rate) follow the *seed-tts-eval* protocol. On ACE-Instruct-TTS-Eval-ZH, CER is computed on 481 utterances, excluding 169 containing dialect speech due to unreliable ASR performance on dialects. **IFMOS** measures human-rated instruction following, reported as mean $\pm$ half-width of 95% bootstrap CI (Appendix A.6).

Implementation. Stage 1 and Stage 2 use a learning rate of 2×10^{-5} and a per-GPU batch size of 4, trained for 2 epochs; Stage 3 uses 5×10^{-6}, one prompt per GPU, and a rollout group of $G = 8$, trained for 200 steps, giving 32 prompts and 256 rollouts per step. Training runs on 32 NVIDIA H20 GPUs across four nodes. Unless otherwise specified, we generate one candidate per input without CPMI-based reranking. We use the same decoding configuration across all ACE-Instruct-TTS experiments; detailed sampling parameters and the inference-time CFG implementation are provided in Appendix A.4. For voice design, no reference audio is provided.

5.2 Results

Main results. Table 2(a) shows that ACE-Instruct-TTS-RL achieves the highest overall IFR of 75.23%, exceeding Qwen-Audio-3.0-TTS-Plus and CosyVoice3 by 2.92 and 8.92 percentage points, respectively. Compared with SFT, SR-GRPO improves IFR by 8.15 points with similar CER (1.97% versus 2.01%). SIM decreases from 0.756 to 0.733 but remains higher than all external baselines, indicating an adherence–speaker-preservation trade-off under the current training configuration. VoxCPM2 achieves the lowest CER (1.22%) but lower IFR and SIM, highlighting distinct quality dimensions. Human ratings show the same trend: the RL model achieves the highest mean IFMOS of 3.90 ± 0.18, compared with 3.64 ± 0.23 for SFT and 3.76 ± 0.24 for Qwen-Audio-3.0-TTS-Plus. All six systems have identical rankings by mean IFMOS and overall IFR, providing complementary evidence for the instruction-following trends (see Appendix A.6 for the human evaluation protocol). Across categories (Table 2(b)), SR-GRPO improves over SFT in all six categories, with the largest gains in acoustic controls (+22.00 percentage points) and style (+15.00 points). The RL model leads in emotion, acoustic controls, compositional conditions, and conflicting conditions, while Qwen-Audio-3.0-TTS-Plus leads in style and dialect.

Table 2: ACE-Instruct-TTS-Eval-ZH results: (a) overall metrics; (b) category-wise IFR. IFR and CER are percentages; IFMOS includes 95% confidence intervals. Qwen and MiMo denote Qwen-Audio-3.0-TTS-Plus and MiMo-V2.5-TTS, respectively. Acous./Comp./Conf. denote acoustic, compositional, and conflicting conditions. **Bold** marks column-best values.

(a) Overall performance

Model	IFR ↑	CER ↓	SIM ↑	IFMOS ↑
CosyVoice3 Du et al. (2025)	66.31	1.90	0.706	3.63 ± 0.24
VoxCPM2 Zhou et al. (2026)	60.15	**1.22**	0.582	3.52 ± 0.25
Qwen Xiang et al. (2026)	72.31	1.66	0.702	3.76 ± 0.24
MiMo Xiaomi (2026)	51.69	2.38	0.669	2.67 ± 0.28
ACE-Instruct-TTS-SFT	67.08	2.01	**0.756**	3.64 ± 0.23
ACE-Instruct-TTS-RL	**75.23**	1.97	0.733	$\mathbf{3.90} \pm 0.18$

(b) Category-wise instruction-following rate

Model	Emotion	Style	Dialect	Acous.	Comp.	Conf.
CosyVoice3	81.11	80.00	68.75	76.00	59.62	30.00
VoxCPM2	68.33	78.00	32.81	82.00	62.18	31.00
Qwen	80.00	**96.00**	**85.94**	70.00	70.51	30.00
MiMo	69.44	63.00	31.25	44.00	51.92	25.00
ACE-Instruct-TTS-SFT	77.22	77.00	79.69	64.00	68.59	30.00
ACE-Instruct-TTS-RL	**83.89**	92.00	81.25	**86.00**	**73.08**	**37.00**

Evaluation on external benchmarks. On the Chinese subset of InstructTTSEval (Table 3), ACE-Instruct-TTS-RL achieves a macro average of 81.2%, outperforming all listed open-source baselines on each of the three tasks. SR-GRPO improves over SFT by 0.7, 2.2, and 3.1 percentage points on APS, DSD, and RP, respectively, raising the macro average by 2.0 points. On SpeechEditBench

Table 3: InstructTTSEval Chinese-subset results Huang et al. (2025) (%, ↑). Avg.: macro average over APS, DSD, and RP. **Bold**: best evaluated system; underline: best among open-source baselines and ours. Reference audio is excluded from highlighting.

Model	APS ↑	DSD ↑	RP ↑	Avg. ↑
Closed-source baselines				
Gemini-2.5-flash-preview-tts	88.2	**90.9**	**77.3**	**85.4**
Gemini-2.5-pro-preview-tts	**89.0**	90.1	75.5	84.8
GPT-4o-mini-TTS	54.9	52.3	46.0	51.1
Open-source baselines				
VoxInstruct	47.5	52.3	42.6	47.5
VoiceSculptor	75.7	64.7	61.5	67.3
Mimo-Audio-7B-Instruct	75.7	74.3	61.5	70.5
FlexiVoice	71.0	71.8	69.7	70.8
Qwen3TTS-12Hz-1.7B-VD	85.2	81.1	65.1	77.1
Ours				
ACE-Instruct-TTS-SFT	86.9	80.9	69.7	79.2
ACE-Instruct-TTS-RL	87.6	83.1	72.8	81.2

Table 4: SpeechEditBench joint success (target success; %, ↑) with selected leading baselines. **Bold**: highest joint success per task. Ours uses re-synthesis with Chinese-translated instructions; baselines use the original protocol, limiting direct comparisons. See Appendix A.7 for all baselines.

Model	Emotion ↑	Style ↑	Prosody ↑
Closed-source baselines			
Gemini-Live	**27.79** (34.43)	63.67 (84.00)	**65.17** (69.67)
GPT-Realtime	14.57 (21.00)	**68.67** (82.33)	63.94 (70.12)
Open-source baselines			
Step-Audio-EditX	7.71 (9.29)	49.67 (54.00)	20.13 (51.51)
VoxCPM2	3.57 (4.14)	32.67 (36.67)	49.00 (50.83)
Ours: adapted protocol			
ACE-Instruct-TTS-SFT	4.64 (5.64)	35.17 (42.00)	47.67 (49.83)
ACE-Instruct-TTS-RL	7.93 (8.86)	35.50 (44.00)	49.83 (52.00)

(Table 4), SR-GRPO improves joint success from 4.64% to 7.93% on Emotion, 35.17% to 35.50% on Style, and 47.67% to 49.83% on Prosody. Target success also improves across all three tasks, although the Style joint-success gain is modest. The RL model is competitive with the listed open-source systems on Emotion and Prosody, but differences in protocol and instruction language limit direct cross-system comparisons. Together, the SFT-to-RL gains across both benchmarks support the effectiveness of SR-GRPO beyond ACE-Instruct-TTS-Eval-ZH.

5.3 Reward Analysis, Ablations, and Cross-Model Validation

All analyses below use ACE-Instruct-TTS-Eval-ZH.

CPMI scoring and candidate selection. Using ACE-Instruct-TTS-SFT, we rank five candidates per input by descending CPMI. IFR decreases monotonically from 71.23% at Rank-0 to 63.08% at Rank-4 (Figure 1(a)), supporting CPMI as an informative within-prompt ranking signal without an external instruction scorer. For Best-of-N selection with $N \in \{1, 2, 4, 8, 16\}$, overall IFR rises from 67.08% at $N = 1$ to 74.31% at $N = 8$, but drops to 71.08% at $N = 16$, showing that larger candidate pools do not guarantee better instruction adherence. With a single candidate and no reranking, SR-GRPO achieves 75.23% overall IFR, exceeding all tested SFT Best-of-N configurations (Figure 1(b)). Relative to Best-of-8, it improves IFR by 0.92 percentage points and SIM by 0.013. These results support policy optimization for improving overall instruction adherence without inference-time reranking. Detailed results, including Conflict-subset trends and evaluation subsets, are provided in Appendix A.5.

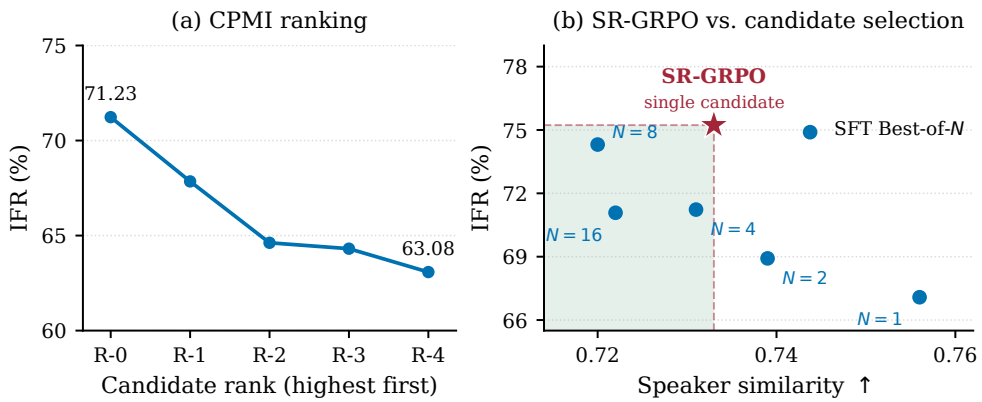

Figure 1: (a) IFR by descending CPMI rank among five SFT candidates per input. (b) Overall IFR vs. SIM for SFT Best-of-N and single-candidate SR-GRPO (star; no reranking). Shading indicates the region dominated by SR-GRPO on both metrics.

Table 5: Reward ablations on ACE-Instruct-TTS-Eval-ZH. **Bold** marks the best value per metric.

Model / Configuration	IFR ↑	CER ↓	SIM ↑
ACE-Instruct-TTS-SFT	67.08	2.01	0.756
ACE-Instruct-TTS-RL: reward formulations			
Ext-RM Binary	67.69	1.93	0.745
Ranked Scalar	69.38	1.87	0.745
Conditional-only	68.77	1.69	0.744
CPMI $(0.6, 0.3, 0.1)$	**75.23**	1.97	0.733
ACE-Instruct-TTS-RL: reward weight ablations			
$(0.0, 0.9, 0.1)$	62.46	2.00	**0.770**
$(0.3, 0.6, 0.1)$	70.31	1.81	0.755
$(0.45, 0.45, 0.1)$	70.00	**1.39**	0.753

Reward ablations. Table 5 compares reward formulations and weights under the same SFT initialization, training data, optimization steps, and decoding settings. With weights $(0.6, 0.3, 0.1)$, *Ext-RM Binary* maps yes/no judgments from Kimi-Audio-7B-Instruct KimiTeam et al. (2025) to $r_{\text{inst}} \in \{1, 0\}$; *Ranked Scalar* assigns rewards from 8 to 1 by descending CPMI within each group, retaining ordering but discarding score gaps; and *Conditional-only* uses $r_{\text{inst}} = -\mathcal{L}_{\text{inst}}$. CPMI achieves the highest IFR (75.23%), exceeding Conditional-only and Ranked Scalar by 6.46 and 5.85 percentage points, respectively. These comparisons support the value of both the instruction-free baseline and score gaps. CPMI also exceeds Ext-RM Binary by 7.54 points, though this difference cannot be attributed solely to binarization because both the scorer and reward format change. Its lower SIM and higher CER than all three alternatives indicate a trade-off rather than uniform improvement across metrics. With $\lambda_{\text{len}} = 0.1$, increasing the timbre weight from $(0.6, 0.3, 0.1)$ to $(0.3, 0.6, 0.1)$ raises SIM from 0.733 to 0.755, nearly matching SFT's 0.756, while lowering IFR from 75.23% to 70.31%. IFR nevertheless remains 3.23 percentage points above SFT, indicating that reweighting recovers much of the speaker similarity while retaining instruction-following gains. Removing the instruction reward, $(0.0, 0.9, 0.1)$, further increases SIM to 0.770, the highest among the evaluated configurations, but reduces IFR to 62.46%, below SFT's 67.08%, with CER of 2.00%. This contrast highlights the importance of the instruction reward for maintaining instruction adherence. Equal instruction and timbre weights, $(0.45, 0.45, 0.1)$, yield the lowest CER (1.39%), with IFR of 70.00% and SIM of 0.753, further illustrating the trade-offs induced by reward weighting.

CFG scaling versus SR-GRPO. Both CFG and SR-GRPO exploit a conditional/unconditional likelihood contrast: CFG applies it at inference time to steer decoding, whereas SR-GRPO uses it as a training reward to update the policy. Raising the guidance weight improves IFR on the SFT

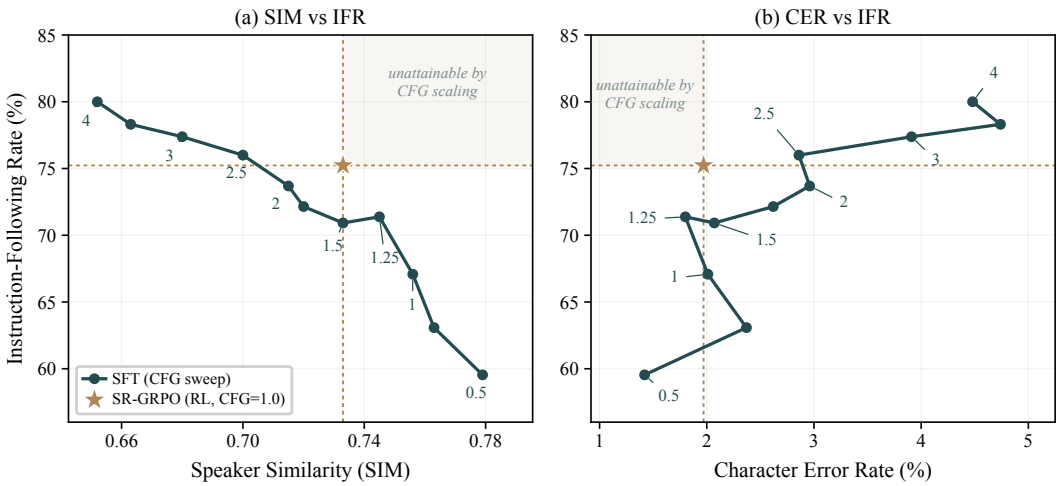

Figure 2: **CFG scaling does not match the SR-GRPO operating point.** IFR vs. speaker similarity (a) and CER (b). Curves: ACE-Instruct-TTS-SFT with CFG 0.5–4.0; star: ACE-Instruct-TTS-RL with CFG 1.0.

Table 6: Cross-model validation of SR-GRPO on ACE-Instruct-TTS-Eval-ZH.

Model	IFR ↑	CER ↓	SIM ↑
CosyVoice3	66.31	1.90	0.706
CosyVoice3 + SR-GRPO	**71.85**	2.00	0.697

model but degrades both speaker similarity and intelligibility (Fig. 2). At CFG=1.25—the highest-IFR configuration retaining SIM≥0.733—SFT reaches 71.38% IFR, which SR-GRPO exceeds by 3.85 points. Matching SR-GRPO on IFR instead requires CFG≈2.5 (76.00%), costing 0.033 SIM and 0.89 points CER. No tested SFT configuration matches or exceeds SR-GRPO on both metrics in either displayed comparison. These results suggest that, under the evaluated decoding settings, SR-GRPO offers a trade-off not obtained by the tested SFT guidance scales.

Cross-model validation. To assess whether the benefits of SR-GRPO extend beyond ACE-Instruct-TTS, we additionally apply it to CosyVoice3. As shown in Table 6, SR-GRPO improves IFR from 66.31% to 71.85%, a gain of 5.54 percentage points. SIM decreases from 0.706 to 0.697, while CER increases from 1.90% to 2.00%. These results provide evidence that the instruction-following benefits extend to a second TTS backbone, while exhibiting an adherence–speaker-preservation trade-off consistent with our ACE-Instruct-TTS experiments. See Appendix A.8 for details.

6 Conclusion

We presented ACE-Instruct-TTS and SR-GRPO for instruction-guided zero-shot speech synthesis. SR-GRPO derives continuous, CPMI-inspired rewards by contrasting instruction-conditioned and instruction-free likelihoods under a frozen SFT model, without an external instruction evaluator. We also introduced ACE-Instruct-TTS-Eval-ZH to jointly assess instruction following, speaker preservation, and intelligibility. SR-GRPO improves IFR from 67.08% to 75.23%, with comparable CER but lower speaker similarity than SFT. Reward ablations support the likelihood contrast and show that reward weighting adjusts this trade-off. Applying SR-GRPO to CosyVoice3 improves IFR by 5.54 percentage points, with a similar speaker-preservation trade-off, providing evidence of applicability to a second TTS backbone. Gains on InstructTTSEval and under an adapted SpeechEdit-Bench protocol provide further evidence that TTS likelihoods offer useful feedback for instruction-following policy optimization. Future work will extend multilingual control and better balance adherence with speaker fidelity.

AI USE DISCLOSURE

We used large language models (LLMs) to assist with drafting, translating, and polishing the manuscript, as well as literature retrieval and discovery, including identifying potentially relevant related work. We used gpt-5.6-luna to generate Chinese instructions and corresponding target texts for ACE-Instruct-TTS-Eval-ZH. The generated instruction–text pairs underwent human review before inclusion in the benchmark.

The authors reviewed the AI-assisted manuscript and verified the cited references. We take responsibility for the final content of this work, including the text, claims, and benchmark materials produced with LLM assistance.

ETHICS STATEMENT

This work aims to improve instruction-guided zero-shot voice cloning, with potential benefits for accessibility and personalized speech generation. However, these capabilities may also be misused for unauthorized voice imitation, impersonation, fraud, and deceptive content creation. Fine-grained control over emotion and speaking style may further enable synthetic speech that misrepresents a reference speaker's attitudes or intentions. Our self-rewarding mechanism optimizes instruction adherence; it does not verify speaker authorization or detect malicious instructions. The reported results evaluate synthesis capabilities and should not be interpreted as evidence of deployment safety.

We train on commercially purchased speech data licensed for model training. These data are not released or redistributed as part of this work. Benchmark reference audio is drawn from seed-tts-eval. Its use and redistribution are subject to applicable licenses, authorization scopes, and privacy requirements; public availability alone does not establish consent for arbitrary voice cloning. The Chinese instructions and corresponding target texts in ACE-Instruct-TTS-Eval-ZH were generated with LLM assistance and reviewed by humans before inclusion. This review assesses benchmark quality and does not replace authorization to use reference voices. Human evaluators participated voluntarily after being informed of the tasks and how their evaluation data would be used, and received compensation for their time.

Our evaluation primarily focuses on Chinese-language settings and does not systematically assess performance disparities across demographic groups. We therefore make no claims about demographic fairness or generalization to untested populations. Before releasing the benchmark, the evaluation code, and our SR-GRPO implementation, we will review redistribution permissions and potential privacy risks. The released code will not include any proprietary training data or model checkpoints; it will support reproduction of the method on publicly available checkpoints and on users' own licensed data. Where redistribution rights for reference audio cannot be confirmed, we will provide source identifiers or access instructions rather than redistribute the recordings. Responsible deployment should require appropriate authorization to use reference voices and clear disclosure that generated audio is synthetic.

REPRODUCIBILITY STATEMENT

We will release the ACE-Instruct-TTS-Eval-ZH benchmark materials and evaluation code, as well as our SR-GRPO implementation for the public CosyVoice3 (with the configuration in Appendix A.8), subject to applicable redistribution permissions and compliance review, so that the core method can be reproduced end-to-end on a public backbone. The commercially purchased training data are licensed for model training but will not be released, and the resulting ACE-Instruct-TTS checkpoints are therefore also withheld; we instead report the data composition, staged training recipe, and all reward and optimization hyperparameters needed to reproduce the method.

REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie

Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models, 2024. URL https://arxiv.org/abs/2406.02430.

Xiong Cai, Dongyang Dai, Zhiyong Wu, Xiang Li, Jingbei Li, and Helen Meng. Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition, 2021. URL https://arxiv.org/abs/2010.13350.

Dekun Chen, Xueyao Zhang, Yuancheng Wang, Kenan Dai, Li Ma, and Zhizheng Wu. Flexivoice: Enabling flexible style control in zero-shot tts with natural language instructions, 2026a. URL https://arxiv.org/abs/2601.04656.

Huakang Chen, Jingbin Hu, Liumeng Xue, Qirui Zhan, Wenhao Li, Guobin Ma, Hanke Xie, Dake Guo, Linhan Ma, Yuepeng Jiang, Bengu Wu, Pengyuan Xie, Chuan Xie, Qiang Zhang, and Lei Xie. Mint-bench: A comprehensive multilingual benchmark for instruction-following text-to-speech, 2026b. URL https://arxiv.org/abs/2604.17958.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. Large-scale self-supervised speech representation learning for automatic speaker verification, 2022. URL https://arxiv.org/abs/2110.05777.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2024. URL https://arxiv.org/abs/2306.05284.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025. URL https://arxiv.org/abs/2505.17589.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right, 2022. URL https://arxiv.org/abs/2104.08315.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report, 2026a. URL https://arxiv.org/abs/2601.15621.

Jingbin Hu, Huakang Chen, Linhan Ma, Dake Guo, Qirui Zhan, Wenhao Li, Haoyu Zhang, Kangxiang Xia, Ziyu Zhang, Wenjie Tian, Chengyou Wang, Jinrui Liang, Shuhan Guo, Zihang Yang, Bengu Wu, Binbin Zhang, Pengcheng Zhu, Pengyuan Xie, Chuan Xie, Qiang Zhang, Jie Liu, and Lei Xie. Voicesculptor: Your voice, designed by you, 2026b. URL https://arxiv.org/abs/2601.10629.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. Instructttseval: Benchmarking complex natural-language instruction following in text-to-speech systems, 2025. URL https://arxiv.org/abs/2506.16381.

Kexin Huang, Liwei Fan, Botian Jiang, Yaozhou Jiang, Qian Tu, Jie Zhu, Yuqian Zhang, Yiwei Zhao, Chenchen Yang, Zhaoye Fei, Shimin Li, Xiaogui Yang, Qinyuan Cheng, and Xipeng Qiu. Moss-voicegenerator: Create realistic voices with natural language descriptions, 2026. URL https://arxiv.org/abs/2603.28086.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang,

Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training, 2023. URL https://arxiv.org/abs/2206.04658.

Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, and Zhizheng Wu. Dualcodec: A low-frame-rate, semantically-enhanced neural audio codec for speech generation, 2025. URL https://arxiv.org/abs/2505.13000.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models, 2016. URL https://arxiv.org/abs/1510.03055.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Chang Liu, Ya-Jun Hu, Ying-Ying Gao, Shi-Lei Zhang, and Zhen-Hua Ling. Group relative policy optimization for text-to-speech with large language models, 2025. URL https://arxiv.org/abs/2509.18798.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance, 2023. URL https://arxiv.org/abs/2306.17806.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. F5r-tts: Improving flow-matching based text-to-speech with group relative policy optimization, 2025. URL https://arxiv.org/abs/2504.02407.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023. URL https://arxiv.org/abs/2301.02111.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis, 2018. URL https://arxiv.org/abs/1803.09017.

Bajian Xiang, Cheng Wen, Han Zhao, Hao Wang, Haoxu Wang, Jiawei Jin, Jiayan Cui, Jie Chen, Mengxi Nie, Tianyu Zhao, Weiqin Li, Xiang Lv, Xiangang Li, Yang Xiang, and Yang Zhou. Qwen-audio-3.0-tts: Freely controllable and highly robust speech synthesis with multi-stage training paradigm, 2026. URL https://arxiv.org/abs/2607.23938.

Xiaomi. MiMo-V2.5-TTS Series: A Family of Speech Synthesis Models Built for the Agent Era. https://mimo.xiaomi.com/mimo-v2-5-tts, April 2026. Accessed: Aug. 27, 2026.

Chee-En Yu, Yi-Cheng Lin, Sung-Feng Huang, Yun-Shao Tsai, Ho-Lam Chung, Xuanjun Chen, and Hung yi Lee. Listen, critique, and refine: Rl-based self-refinement for instruction-following speech synthesis, 2026. URL https://arxiv.org/abs/2609.24163.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2025. URL https://arxiv.org/abs/2401.10020.

Hanlin Zhang, Daxin Tan, Dehua Tao, Xiao Chen, Haochen Tan, and Linqi Song. Speecheditbench: A bilingual multi-attribute benchmark for instruction-guided speech editing, 2026. URL https://arxiv.org/abs/2606.01804.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 554–563. ACM, October 2024. doi: 10.1145/3664647.3681680. URL http://dx.doi.org/10.1145/3664647.3681680.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, Jiaheng Wu, Ziyang Wang, Xudong Shen, Runchuan Ye, Zhisheng Zhang, Jiuyang Zhou, Bingsong Bai, Weiyue Sun, Mengyuan Deng, Qundong Shi, Zhiyong Wu, and Zhiyuan Liu. Voxcpm2 technical report, 2026. URL https://arxiv.org/abs/2606.06928.

A Appendix

A.1 Construction Details of ACE-Instruct-TTS-Eval-ZH

This subsection describes reference selection, candidate generation, and manual verification for ACE-Instruct-TTS-Eval-ZH. The final benchmark contains 650 samples, with category counts reported in Table 7.

Table 7: Composition of ACE-Instruct-TTS-Eval-ZH. Counts denote the number of evaluation samples in each category. Regional dialects and accents are illustrated with representative examples.

Category	Count	Coverage
Emotion & Tone	180	Seven broad categories with finer-grained subtypes: neutral (calm/rational, gentle/empathetic); happiness (joyful, confident/optimistic, affectionate/sociable); anger (irritated/dissatisfied, resentful/humiliated); sadness (grieving, dejected/despairing); surprise (confused, astonished); fear (worried/uneasy, frightened/threatened); and disgust (averse, loathing/rejecting).
Speaking Style	100	Four register groups with eight substyles: interactive (casual conversation, affectionate interaction); informational (formal broadcasting, professional explanation); narrative (lyrical narration, storytelling); and performative (promotional/persuasive delivery, playful/lively delivery).
Chinese Dialects & Accents	64	Regional dialect and accent requests, such as Beijing, Northeastern, Sichuan/Chengdu, Shanghainese, Cantonese, Hunan, Jiangxi/Nanchang, Henan, Tianjin, Shaanxi/Xi'an, Taiwan-associated varieties or accents, Southern Min/Quanzhou, Hakka, Shandong, and Lanzhou.
Acoustic Controls	50	Speaking rate (slow, normal, fast); pitch (low, medium, high); volume (low, medium, high); and whispering as a phonation-mode control.
Compositional Conditions	156	Instructions combining two, three, or four attributes, such as "whispering sadly at a slow pace."
Conflicting Conditions	100	Instruction–text conflicts, such as cheerful delivery of sad content; and instruction–reference conflicts, such as a shouting instruction paired with a calmly spoken reference.
Total	**650**	

Reference selection and candidate generation. Reference clips and their transcripts are randomly sampled from *seed-tts-eval* Anastassiou et al. (2024), and gpt-5.6-luna generates the paired instructions and target texts. The prompt specifies an attribute inventory and six condition types, covering single-attribute control, multi-attribute composition, and contextual conflicts, and requests ten candidates per call without fixing the final category distribution.

Verification and prompt. All samples are manually inspected and rewritten when unsuitable; for instruction–reference conflicts, the selected recording is checked against the requested delivery, since the generated `reference_audio_desc` field states an intended condition rather than the actual one. Conflict cases thus test instruction following under incongruent textual or reference cues, not the satisfaction of self-contradictory instructions. The Chinese prompt is reproduced below in two panels (attributes and task design; constraints and output examples), with formatting adapted and decorative symbols omitted.

Generation Prompt / 01: Attributes and Task Design

角色设定

你是一个专业的语音合成（TTS）数据构建专家。 你的任务是根据我提供的【属性标签库】与【6大类指令规则】， 生成用于测试零样本可控 TTS 模型的“自然语言控制指令（Instruction）”以及“对应的朗读文本（Text）”。

【属性标签库】（请严格从中抽取）

1. 情感 **(Emotion)**

中性：平静理性、温柔共情； 快乐：愉悦喜悦、自信乐观、亲昵社交； 愤怒：烦躁不满、怨恨受辱； 悲伤：痛苦哀伤、失落绝望； 惊讶：困惑冲击、惊愕震撼； 恐惧：担忧不安、恐惧威胁； 厌恶：嫌弃反感、憎恶抗拒。

2. 风格 **(Style)**

交互性语域：日常闲聊、社交亲昵； 信息性语域：官方演播、专业解说； 叙事性语域：旁白抒情、故事讲述； 表现性语域：广告煽动、趣味活泼。

3. 方言 **(Dialect)**

泛指：方言、口音。

官话变体：北京话/腔、东北话/味、四川话/成都话、 天津话、陕西话/西安话、山东话/济南话/青岛话、兰州话。

其他方言：上海话/腔、广东话/粤语、湖南话、 江西话/南昌话、河南话、台湾话/腔、 闽南话/泉州话、客家话。

4. 声学参数 **(Parameters)**

语速：慢、正常、快； 音高：耳语、低沉、适中、高亢； 音量：低、适中、高。

【生成任务要求】

请生成 **10**个 测试用例， 必须严格涵盖以下 **6**大类指令：

【单一属性控制】（共 **4** 个）

类别 **1**：单一情感 **(1**个**)**： 指令中仅包含 1 种情感要求。

类别 **2**：单一风格 **(1**个**)**： 指令中仅包含 1 种风格要求。

类别 **3**：单一方言 **(1**个**)**： 指令中仅包含 1 种方言要求。

类别 **4**：单一参数 **(1**个**)**： 指令中仅包含 1 种声学参数（语速/音高/音量）要求。

【复杂控制】（共 **6** 个）

类别 **5**：多属性组合 **(4**个**)**： 随机组合 2个、3个 或 4个 属性 （例如：情感+方言，或 风格+语速+音量+情感）。

类别 **6**：冲突测试 **(2**个**)**：

指令-文本冲突 **(1**个**)**： 指令要求的情感/风格，与文本本身的语义完全相反 （例如：用极度愉悦喜悦的声音， 播报一段悲惨的灾难新闻）。

指令-参考音频冲突 **(1**个**)**： 设定一个参考音频的基调，但指令要求生成完全相反的输出。 请在输出中增加 `reference_audio_desc` 字段描述参考音频。

Generation Prompt / 02: Constraints and Output Examples

指令与文本编写要求

指令口语化： 将选中的属性融合成自然流畅的口语化指令， 绝不能生硬堆砌标签。 例如，不要说“使用恐惧、东北话、快语速”， 而应该说“用东北口音，语速快一点， 表现出仿佛看到鬼一样的恐惧感”。

文本匹配： 除了“指令-文本冲突”类别外， 其余类别的文本内容需与指令的语境高度匹配。

【严格限制】（非常重要）

数字正则化： 生成的“朗读文本（Text）”中， 绝对不允许出现任何阿拉伯数字（**0-9**）。 所有的数字、年份、金额、时间等， 必须全部转换为中文汉字。

错误示范：2023年，150块，3.14。

正确示范：二零二三年，一百五十块，三点一四。

【输出格式】

请使用以下 JSON 格式输出结果。

```
[
{
  "id": 1,
  "type": "类别1-单一情感",
  "instruction":  "用充满恐惧的情绪说，仿佛看到了鬼一样。",
  "text":  "那个黑影刚才明明还在窗外，怎么一眨眼的功夫， 就站到我床头了？",
  "reference_audio_desc": "无"
},
{
  "id": 5,
  "type": "类别5-多属性组合(3属性)",
  "instruction":  "用东北话说，语速搞快点，语气要显得特别烦躁和不满。",
  "text":  "这都下午三点半了，公交车怎么还没来？ 我等了整整四十五分钟了，冻死我了！",
  "reference_audio_desc": "无"
},
{
  "id": 10,
  "type": "类别6-冲突(指令参考冲突)",
  "instruction":  "我用最高亢的音量，愤怒地吼出来！",
  "text":  "你凭什么动我的东西？我跟你说过一百遍了， 别碰我的电脑！",
  "reference_audio_desc":  "一段情绪非常平静理性的男声朗读。"
}
]
```

A.2 Architecture Diagram of ACE-Instruct-TTS

Figure 3 illustrates the architecture described in Section 3.1, including instruction and speaker conditioning, multi-channel RVQ token prediction, and waveform reconstruction.

A.3 Objective Details

Masking and weighting. $m_{i,s,j}$ restricts the sums in Eq. equation 4 to generated cells: the conditioning prefix is excluded, the per-channel delay offset is applied so that channel j starts j frames later than channel 1, and all positions at or beyond the EOS of channel 1 are masked. The channel weights $w_j \propto (12, 8, 6, 4, 2, 2, 1, 1)$ are rescaled to unit mean, so $Z_i = \sum_{s,j} w_j m_{i,s,j}$ is on the scale of the number of generated cells. The objective and the KL penalty use the same $m_{i,s,j}$, the same

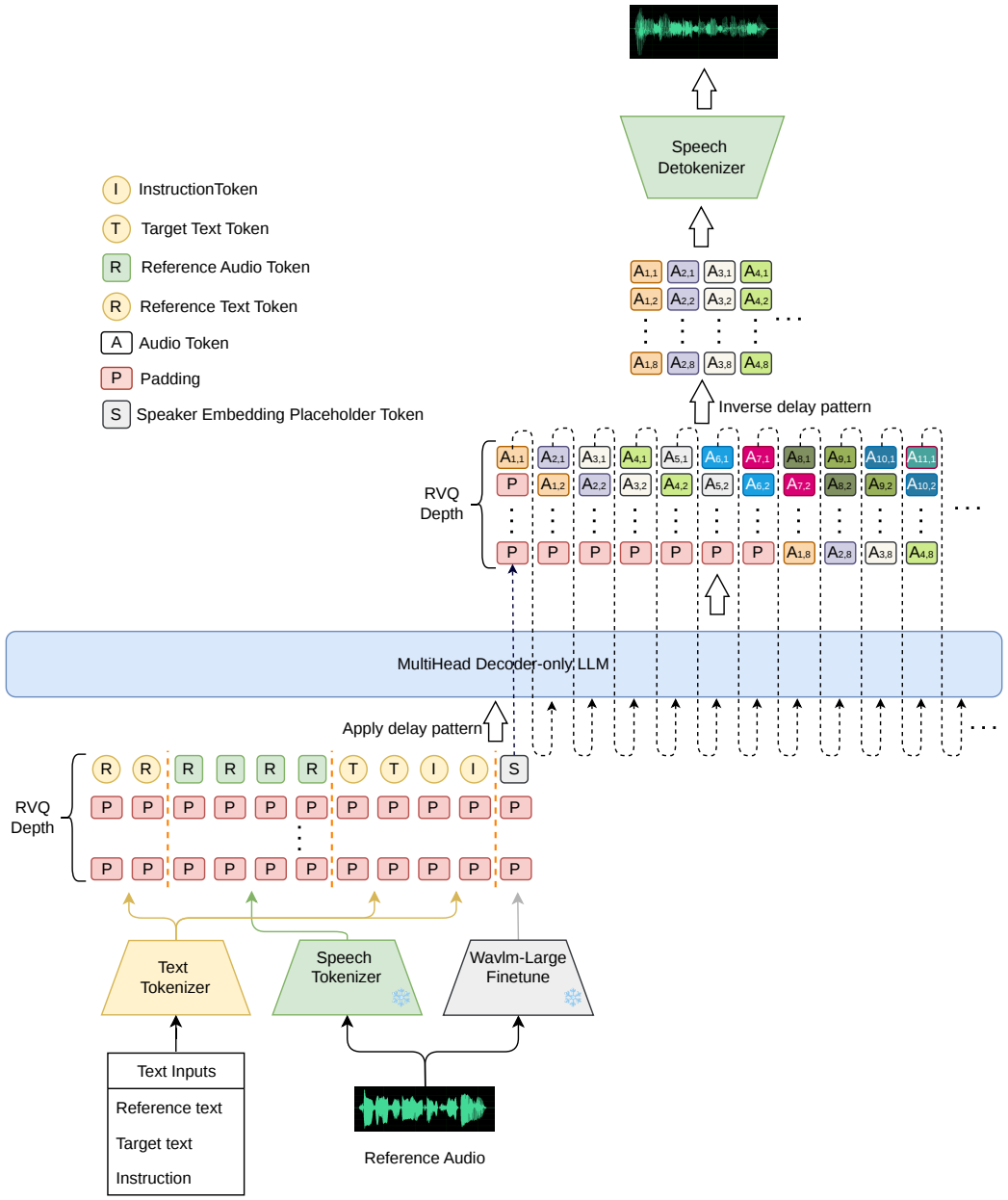

Figure 3: Architecture of ACE-Instruct-TTS. The decoder-only backbone is conditioned on the target text, a natural-language instruction, and a reference speaker embedding. The instruction text is inserted between the target text and the speaker embedding placeholder. For voice design, no reference audio is provided, and the embedding matrix of the Speaker Embedding Placeholder remains all zeros. The model predicts eight RVQ token channels using a delay pattern, and the waveform is reconstructed using a flow-matching model followed by BigVGAN-v2.

w_j and the same Z_i, so the two terms are commensurable and the KL target below is expressed in nats per w-weighted cell.

KL estimator. $\hat{k}^{(3)}_{i,s,j} = e^u - u - 1$ with $u = \log \pi_{\text{ref}} - \log \pi_\theta$ evaluated per cell, with u clipped to $[-10, 10]$ and $\hat{k}^{(3)}$ to $[0, 5]$ for numerical stability.

Adaptive KL coefficient. Let D_t denote the value of the KL term of Eq. equation 4 at step t, i.e. exactly the penalized quantity rather than a separate diagnostic. The coefficient follows a proportional controller in the *relative* error, which is clipped before being applied:

$$\beta_{t+1} = \mathrm{clip}\Big(\beta_t\Big(1 + K\,\mathrm{clip}\big(\tfrac{D_t - D^\star}{D^\star}, -\delta, +\delta\big)\Big),\ \beta_{\min},\ \beta_{\max}\Big), \tag{5}$$

with $D^\star = 0.3$, $K = 0.5$, $\delta = 0.2$, $\beta_0 = 0.01$ and $[\beta_{\min}, \beta_{\max}] = [0.01, 1.0]$. The inner clip bounds the per-step multiplier to $[1 - K\delta, 1 + K\delta] = [0.9, 1.1]$. Gradient accumulation is 1, so each step consumes one fresh group and the controller runs once per step, on averaged over all 256 rollouts of the step, immediately after the parameter update; there is no exponential moving average, no warm-up and no freeze period. The loss at step t therefore uses β_t carried over from step $t - 1$, and the controller produces β_{t+1}.

A.4 Decoding Settings

Shared decoding configuration. All experiments with ACE-Instruct-TTS use stochastic sampling with temperature 1.0, top-k filtering with $k = 50$, and nucleus sampling with $p = 0.95$. The repetition penalty is 1.0, corresponding to no repetition adjustment. These settings are shared across all eight RVQ channels, the SFT and RL models, and the candidate-scoring, reward-ablation, and CFG experiments. The maximum generation budget is 5,000 delayed decoding steps, with up to eight channel tokens predicted at each step. Generation follows the model's delay-pattern masking and termination rules. Unless otherwise specified, we generate one candidate per input without reranking; candidate-scoring experiments use the sampling budgets reported in Appendix A.5. We use Transformers version 4.52.1.

Inference-time instruction CFG. For CFG experiments, we run instruction-conditioned and instruction-free branches of the same model, retaining the target text and speaker conditioning in both branches. At decoding step t and RVQ channel j, the pre-softmax logits are combined as

$$\ell_{t,j}^{(s)} = \ell_{t,j}^{\mathrm{no-inst}} + s\left(\ell_{t,j}^{\mathrm{inst}} - \ell_{t,j}^{\mathrm{no-inst}}\right), \tag{6}$$

where s denotes the guidance scale. The branches maintain separate KV caches and receive identical generated speech tokens. The shorter prefix is left-padded with masked tokens, and predictions are aligned by decoding step and RVQ channel. We apply delay-pattern token masks to the combined logits, followed by temperature scaling, top-k filtering, and top-p filtering using the shared configuration above. No additional CFG-based candidate filtering is used. We evaluate guidance scales from 0.5 to 4.0 on the SFT model and compare against the RL model at $s = 1$. At $s = 1$, decoding directly uses the standard instruction-conditioned generation routine.

A.5 Numerical Results for CPMI-Based Candidate Scoring

Table 8 reports the numerical results for CPMI-based candidate scoring and selection underlying Figure 1. All experiments use a fixed ACE-Instruct-TTS-SFT policy and cover all 650 inputs in ACE-Instruct-TTS-Eval-ZH.

For Table 8(a), we generate five candidates per input and rank them by descending CPMI. Candidates at each rank form a 650-utterance evaluation set, with Rank-0 containing the highest-scoring candidates. IFR decreases monotonically from 71.23% at Rank-0 to 63.08% at Rank-4, supporting CPMI as an informative within-prompt ranking signal.

Table 8(b) reports IFR and SIM for Best-of-N selection under the evaluation protocol in Section 5.1. For each input, we select the highest-CPMI candidate among $N \in \{1, 2, 4, 8, 16\}$ generated candidates. Overall IFR increases from 67.08% at $N = 1$ to 74.31% at $N = 8$, while SIM decreases from 0.756 to 0.720. At $N = 16$, overall IFR drops to 71.08%. By contrast, IFR on the 100-input Conflict subset is non-decreasing across the tested budgets, rising from 30.00% to 44.00%. These results support CPMI as a useful candidate-selection signal, while showing that larger sampling budgets do not consistently improve overall adherence or preserve speaker similarity.

Table 8: CPMI-based candidate scoring with ACE-Instruct-TTS-SFT. (a) IFR by descending CPMI rank among five candidates per input; Rank-0 denotes the highest-scoring candidate. (b) Best-of-N selection using CPMI. Overall IFR and SIM use all 650 inputs; Conflict IFR uses the 100 conflicting-condition inputs. IFR is reported as a percentage. **Bold** marks the highest value in each row.

(a) Instruction following by CPMI rank

Subset	Rank-0	Rank-1	Rank-2	Rank-3	Rank-4
Overall	**71.23**	67.85	64.62	64.31	63.08

(b) Best-of-N candidate selection

Subset	Metric	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 16$
Overall	IFR ↑	67.08	68.92	71.23	**74.31**	71.08
	SIM ↑	**0.756**	0.739	0.731	0.720	0.722
Conflict	IFR ↑	30.00	36.00	36.00	41.00	**44.00**

A.6 Human Evaluation Setup

Evaluation setup. We assess perceived instruction adherence using instruction-following mean opinion scores (IFMOS). We construct a category-balanced subset of ACE-Instruct-TTS-Eval-ZH by randomly selecting five inputs from each of its six instruction categories: emotion, style, dialect, acoustic controls, compositional conditions, and conflicting conditions. The resulting subset contains 30 inputs.

We evaluate six systems on the same inputs: CosyVoice3, VoxCPM2, Qwen-Audio-3.0-TTS-Plus, MiMo-V2.5-TTS, ACE-Instruct-TTS-SFT, and ACE-Instruct-TTS-RL. This yields 180 generated audio clips. Nine trained raters each evaluate all 180 clips, providing nine ratings per clip and 1,620 IFMOS ratings in total.

Double-blind evaluation. Neither the raters nor the experimenters administering the listening test know the system identity associated with each audio clip during the rating process. Audio clips from all six systems are pooled, randomly shuffled, and renamed using anonymous sample identifiers. The mapping between anonymous identifiers and system identities is withheld from both groups throughout the rating process. System identities are recovered for statistical aggregation after rating is complete. The shuffling procedure uses a fixed random seed of 20260904 for reproducibility.

Each rater evaluates the clips individually in randomized order rather than in system-specific groups. For each item, the rater reads the corresponding natural-language instruction and listens to the generated audio identified by its anonymous sample ID. For IFMOS, only the generated audio is assessed; listening to reference audio is not required.

Scoring objective and rubric. IFMOS measures how accurately and completely synthesized speech follows a natural-language instruction. Raters assess whether requested attributes, such as emotion, speaking style, dialect, speaking rate, and volume, are appropriately realized. They are instructed not to base their scores on general audio quality or speaker timbre. Vocal attributes explicitly requested by the instruction remain part of the instruction-following assessment.

Each clip receives an integer score from 1 to 5 according to the rubric in Table 9. Higher scores indicate more accurate and complete instruction following.

Guidelines for different instruction types. Raters apply the shared scoring rubric with the following guidance:

- **Single-dimension instructions.** Assess whether the requested attribute, such as emotion, style, dialect, or speaking rate, is correctly expressed at an appropriate intensity.
- **Compositional conditions.** Assess whether all requested dimensions are satisfied jointly. For example, an instruction to "whisper slowly and sadly" requires whispering, a slow

Table 9: Five-point rating rubric for instruction-following MOS (IFMOS). Scores reflect adherence to the instruction, rather than general audio quality or speaker timbre.

Score	Rating	Criterion
5	Perfect	Fully and precisely satisfies all requirements of the instruction, with the requested expressive attributes clearly and sufficiently realized.
4	Good	Satisfies the main requirements with appropriate expression, with only minor shortcomings in details or expressive intensity.
3	Partial	Follows only part of the instruction, such as satisfying only some dimensions of a multidimensional request, or expresses the requested attributes with clearly insufficient intensity.
2	Weak	Shows only weak signs of the requested attributes, with most instruction requirements not realized.
1	Ignored	Does not follow the instruction; the speech is unrelated to or contrary to the requested attributes.

speaking rate, and sadness. Satisfying only a subset of these requirements constitutes partial instruction following.

- **conflicting conditions.** Assess whether competing requirements are reasonably resolved in accordance with the instruction, respecting any explicitly stated priorities. Instruction adherence takes precedence over a preference for a more natural or pleasant rendition.

Score aggregation. We first average the nine ratings assigned to each audio clip to obtain a clip-level mean score. The system-level IFMOS is then computed as the arithmetic mean of the 30 clip-level scores for that system. Because each category contains five inputs and every clip receives nine ratings, this aggregation assigns equal weight to each input and each category.

Inter-rater reliability. All nine raters provided ratings for all 180 audio clips. We assessed inter-rater reliability using a two-way random-effects, absolute-agreement intraclass correlation coefficient (ICC). Single-rater reliability was moderate ($\mathrm{ICC}(2,1) = 0.583$), while reliability of the nine-rater mean used in our IFMOS analysis was high ($\mathrm{ICC}(2,9) = 0.926$). In a leave-one-rater-out analysis, single-rater absolute-agreement ICC ranged from 0.555 to 0.610, suggesting that the overall reliability estimate was not driven by any single rater. These estimates describe rating reliability across the pooled clips from all six systems.

Confidence intervals. We estimate 95% confidence intervals using a nonparametric percentile bootstrap with 10,000 resamples and a fixed random seed of 20260904. For each system, each bootstrap replicate samples 30 clip-level mean scores with replacement from the observed 30 scores and computes their mean. The 2.5th and 97.5th percentiles of the bootstrap means define the lower and upper confidence limits.

Resampling is performed over inputs after averaging the nine ratings for each clip, rather than treating the 270 individual ratings per system as independent observations. The bootstrap is not stratified by instruction category, and raters are not separately resampled. The resulting intervals therefore characterize input-sampling variability conditional on the participating rater panel.

Reporting convention. For compact presentation, Table 2 reports the observed mean $\pm$ half the width of its 95% bootstrap confidence interval. Since percentile bootstrap intervals can be asymmetric about the observed mean, this notation summarizes interval width rather than specifying the exact lower and upper confidence limits.

IFMOS is computed on the 30-input human-evaluation subset. Automatic IFR and SIM are computed on all 650 inputs, whereas CER is computed on the 481-input subset specified in Section 5.1.

A.7 Full Results on SpeechEditBench

Table 10 expands Table 4 with all baseline systems considered in our comparison. Baseline results are taken from the published benchmark. Our models use an adapted reference-conditioned resynthesis protocol with instructions translated into Chinese, while retaining the original transcript

and output languages. These protocol differences limit direct cross-system comparisons; the SFT-to-RL comparison uses the same adapted protocol.

Table 10: Full SpeechEditBench results: joint success (target success; %, ↑). **Bold**: highest reported joint success per task. Ours uses an adapted re-synthesis protocol; baselines use the original benchmark protocol.

Model	Emotion ↑	Style ↑	Prosody ↑
Closed-source systems			
Gemini-Live	**27.79** (34.43)	63.67 (84.00)	**65.17** (69.67)
GPT-Realtime	14.57 (21.00)	**68.67** (82.33)	63.94 (70.12)
Open-source systems			
Ming-UniAudio	3.43 (5.29)	22.17 (32.50)	26.50 (28.00)
Step-Audio-EditX	7.71 (9.29)	49.67 (54.00)	20.13 (51.51)
Qwen3-Omni	1.64 (15.29)	24.17 (63.50)	38.17 (44.83)
Kimi-Audio	2.36 (22.43)	24.50 (63.33)	13.50 (40.33)
Mimo-Audio-Base	0.21 (8.79)	5.83 (42.17)	5.67 (36.17)
Mimo-Audio-Instruction	0.86 (42.36)	0.50 (77.50)	0.67 (42.83)
VoxCPM2	3.57 (4.14)	32.67 (36.67)	49.00 (50.83)
Ours: adapted protocol			
ACE-Instruct-TTS-SFT	4.64 (5.64)	35.17 (42.00)	47.67 (49.83)
ACE-Instruct-TTS-RL	7.93 (8.86)	35.50 (44.00)	49.83 (52.00)

A.8 Cross-Model Validation Details: CosyVoice3

Setup. We apply SR-GRPO to the public CosyVoice3 LLM checkpoint (`Fun-CosyVoice3-0.5B` backbone), updating only the autoregressive speech language model; the flow-matching decoder and vocoder remain frozen. The frozen scorer π_{ref} is initialized from the *same* checkpoint as the policy and is never updated, so policy and scorer coincide at step 0 and diverge only through policy optimization.

Unlike our model, CosyVoice3 predicts a *single* stream of semantic tokens: speaker identity is not encoded in the LLM output but is injected downstream, by conditioning the flow-matching decoder on a speaker embedding. We consequently disable this term for CosyVoice3 and optimize the instruction and duration rewards only.

Training configuration. Table 11 lists the full hyperparameter setting.

Table 11: SR-GRPO configuration for CosyVoice3.

Learning rate / precision	5×10^{-6} / FP32
GPUs	8
Batch × accumulation	1×1 per GPU
Group size G	8
Sampling	top-k 25, top-p 0.8
Reward weights	instruction 0.9, duration 0.1
Speaker-similarity reward	disabled
Adaptive KL	target 0.3, coef. $\in [0.01, 1.0]$
Training steps	200

Results. Table 12 reports instruction-following rate (IFR), speaker similarity (SIM), and character error rate (CER) before and after SR-GRPO. The $+5.54$ point IFR gain is obtained after only 200 optimization steps, without any backbone-specific tuning of the reward.

Table 12: Cross-model validation on CosyVoice3. Best IFR in bold.

	IFR (%) ↑	SIM ↑	CER (%) ↓
CosyVoice3	66.31	0.706	1.90
+ SR-GRPO (200 steps)	**71.85**	0.697	2.00
Δ	$+5.54$	-0.009	$+0.10$