CycleSpeech: Reciprocal Alignment for Instruction-Controlled Speech Synthesis and Paralinguistic Understanding

Anonymous authors
Paper under double-blind review

Abstract

Instruction-controlled speech synthesis and paralinguistic understanding are often trained independently, leaving reciprocal feedback between the two tasks underexplored. We introduce **CycleSpeech**, a framework that connects generation and understanding through a shared, structured *voice profile* that serves as a common target for supervision and reciprocal feedback. The forward cycle assesses whether synthesized speech expresses the intended attributes by comparing recovered and target profiles. The backward cycle evaluates whether profiles inferred from real speech can guide reconstruction of the source speaking style. To support both directions, we construct a bilingual dataset of 20,046 examples pairing instructions, target speech, speaker references, and structured profiles. Building on joint supervised fine-tuning, **CycleGRPO** alternates policy updates using reciprocal rewards grounded in profile consistency and speaking-style reconstruction. Fixed target profiles anchor feedback from the evolving counterpart. This procedure requires neither human preference annotations nor an additional preference-trained reward model. Evaluations on Chinese and English benchmarks show improved instruction adherence and profile recovery while maintaining competitive synthesis quality. Compared with Step-Audio-2-mini, CycleSpeech improves instruction-match accuracy by 4.50 percentage points in Chinese and 10.06 in English. The corresponding gains in the overall speech-to-profile score are 21.36 and 17.39 points. Controlled ablations further support the contribution of cycle feedback to generation control. These results support structured voice profiles as an interface for reciprocal training between speech generation and paralinguistic understanding. An online demo is available at `https://cyclespeech.github.io`.

1 Introduction

Instruction-controlled text-to-speech (TTS) aims to realize user-specified vocal attributes while preserving naturalness and intelligibility. Text-based control has expanded from predefined style descriptions Guo et al. (2023); Yang et al. (2024); Ji et al. (2024a) to open-ended instructions Ren et al. (2026); Zhou et al. (2024); Yang et al. (2025), sometimes combined with speaker references Chen et al. (2026); Ji et al. (2024b). Yet natural-sounding speech can still miss the requested emotion, vocal texture, or delivery. The challenge is therefore not only to interpret an instruction, but also to verify that its attributes are audible in the output.

Recent approaches make the intended vocal delivery explicit before synthesis. OV-InstructTTS Ren et al. (2026) infers emotion labels, acoustic descriptions, and paralinguistic tags before synthesis. BatonVoice Wang et al. (2026) uses an external language model to plan vocal features for a separate TTS model. Such planning specifies the intended delivery but does not verify its realization in synthesized speech. Conversely, a description inferred from speech may match an annotation without preserving enough information to reproduce the source's speaking style.

We therefore ask: *Can speech generation and paralinguistic understanding provide useful training feedback to each other?* We introduce **CycleSpeech**, which uses a structured *voice profile* as a common target for understanding supervision and generation verification. The profile describes speaker traits, prosody, affect, and context in a canonical schema, making recovered attributes comparable

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.

with their targets. Generation still takes natural-language instructions as input; the target profile provides training supervision rather than an additional inference-time requirement.

The shared profile connects the tasks through two complementary training checks. In the forward loop, understanding recovers a profile from generated speech and compares it with the target, supplying feedback on realized control. In the backward loop, generation reconstructs speech from an inferred profile, testing whether the description preserves the source's speaking style. Together, these loops evaluate profiles through both agreement with annotations and their consequences for synthesis. Figure 1 contrasts this reciprocal design with separate and parameter-shared systems.

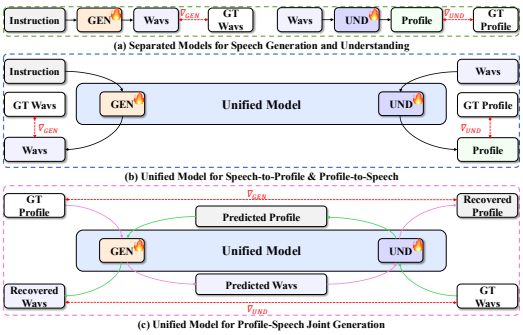

Figure 1: Generation–understanding paradigms. (a) Separate models optimize generation and understanding independently. (b) A shared backbone supports both tasks but still receives one-way supervision. (c) CycleSpeech closes reciprocal profile–speech cycles, allowing generation and understanding to verify each other.

Reciprocal training requires a shared representation linking generation targets to attributes inferred from speech. Existing resources pair speech with structured style factors or natural-language captions, and some support both TTS and style captioning Ji et al. (2024a); Kawamura et al. (2024); Jin et al. (2024); Diwan et al. (2025); Wang et al. (2025); Zhu et al. (2024), but lack a canonical profile that jointly supports understanding supervision, field-level generation rewards, and reciprocal optimization. We construct 20k bilingual examples that align each utterance with its transcript, speaker reference, instruction, and canonical voice profile. This profile provides a reference-grounded interface across both directions: it supervises speech-to-profile prediction, provides field-level rewards for generated speech, and conditions speaking-style reconstruction.

We first establish a compatible initialization through joint supervised fine-tuning (Joint SFT), adapting understanding to generated speech after a generation warm-up. We introduce **CycleGRPO**, which alternately optimizes generation and understanding through reciprocal feedback. The understanding branch provides unified profile-verification feedback without separate attribute-specific reward classifiers, while training requires neither human preference annotations nor an additional preference-trained reward model. Our contributions are:

- We introduce CycleSpeech, connecting instruction-controlled speech synthesis and paralinguistic understanding through a shared, recoverable voice profile. The forward cycle verifies whether generated speech realizes the intended attributes, while the backward cycle tests whether inferred profiles support reconstruction of the source speaking style.
- We develop CycleGRPO to optimize both cycles through alternating policy updates, with evolving counterparts providing feedback and fixed profile targets anchoring learning. It preserves natural-language generation inputs and requires neither human preference annotations nor a separately trained preference reward model.
- We construct a bilingual instruction–speech–profile dataset that jointly supports understanding supervision and generation verification. Bilingual evaluations and controlled ablations demonstrate the importance of reciprocal cycle feedback, with improved instruction following and categorical profile recovery over no-cycle and fixed-feedback baselines.

2 Related Works

2.1 Instruction-Controlled Speech Generation

PromptTTS Guo et al. (2023), InstructTTS Yang et al. (2024), Parler-TTS Lyth & King (2024), and Audiobox Vyas et al. (2023) condition synthesis on textual descriptions. ControlSpeech Ji et al. (2024b) and VoxInstruct Zhou et al. (2024) combine instruction control with zero-shot voice cloning. Recent methods add open-vocabulary reasoning, vocal-feature planning, or joint style–

timbre control Ren et al. (2026); Wang et al. (2026); Zhang et al. (2025b). Post-training further refines synthesis through task-specific rewards: CosyVoice 3 Du et al. (2025) uses ASR rewards, while Fish Audio S2 Liao et al. (2026) uses ASR, quality, and speaker feedback. FlexiVoice Chen et al. (2026) progresses from emotion-focused DPO and GRPO to complex instruction following guided by audio-language-model rewards. These methods optimize synthesis, whereas CycleSpeech jointly aligns generation and paralinguistic understanding through profile verification and speaking-style reconstruction. Attribute- and caption-based datasets Ji et al. (2024a;b); Jin et al. (2024); Diwan et al. (2025); Wang et al. (2025) provide descriptive supervision. Our corpus further supplies canonical profiles as shared targets for understanding supervision and generation verification.

2.2 Verifiable Shared Targets for Speech Generation and Understanding

Connecting generation and understanding requires a shared target that supports both generation and verification. Structured descriptions provide such an interface by mapping instructions to explicit paralinguistic attributes that can be recovered from speech. SpeakerCard-1M Peng et al. (2026) organizes speaker and utterance traits into a typed schema, while UniStyle Zhu et al. (2024) supports both speaking-style captioning and synthesis through shared text descriptions. However, its two directions remain independently supervised. CycleSpeech instead uses a structured voice profile as an auditable interface for understanding and generation, making reciprocal verification well defined.

2.3 Paralinguistic Speech Understanding and Evaluation

External evaluators and understanding models can judge or describe speaking style. InstructTTSEval Huang et al. (2025) and audio-aware LLM judges Chiang et al. (2025) assess instruction–speech consistency, whereas StyleCap Yamauchi et al. (2024) and ParaSpeechCLAP Diwan et al. (2026) caption or align broader style semantics. Speech-chain methods jointly train ASR and TTS by reconstructing speech, text, or intermediate representations Tjandra et al. (2017); Hori et al. (2019). Cycle consistency has also been used to preserve reference styles during TTS style transfer Whitehill et al. (2020); Xue et al. (2021). Other work introduces speaker-consistency constraints and staged optimization to prevent TTS from overfitting to the ASR objective Makishima et al. (2022). CycleSpeech extends this principle to an explicit, recoverable voice profile linking instruction-controlled synthesis and paralinguistic understanding. Attribute verification and speaking-style reconstruction supply reciprocal feedback for alternating policy updates in CycleGRPO, while generation retains its natural-language instruction interface.

3 Method

3.1 Overview

CycleSpeech couples instruction-controlled generation and paralinguistic understanding through a structured voice profile (Figure 2). GEN takes an instruction, transcript, and speaker reference waveform as inputs. The gold profile provides supervision and verification targets, rather than an additional conditioning input to GEN. A frozen Step-Audio-2-mini Wu et al. (2025) backbone $\mathcal{M}_\phi$ is shared by two LoRA adapters: GEN, parameterized by θ_G, and UND, parameterized by θ_U. Training comprises Joint SFT for compatible bidirectional initialization and CycleGRPO for reciprocal alignment. The waveform and profile interfaces are discrete, so only the active adapter receives gradients in each optimization direction.

3.2 Training Data Construction

Each training instance contains an instruction I, transcript T, speaker reference R, target speech x^* with discrete tokens A^*, and canonical voice profile P^*. The same profile serves as the supervision target for UND and the verification target for GEN, enabling bidirectional training through aligned annotations. Following InstructTTSEval Huang et al. (2025), instructions span three categories: Acoustic-Parameter Specification (APS) for acoustic control, Descriptive-Style Directive (DSD) for high-level style descriptions, and Role-Play (RP) for contextual scenarios and character traits.

Profile and prompt construction. We curate human speech from WenetSpeech, a multi-domain Mandarin corpus spanning diverse speaking styles, and an internal collection of English film and

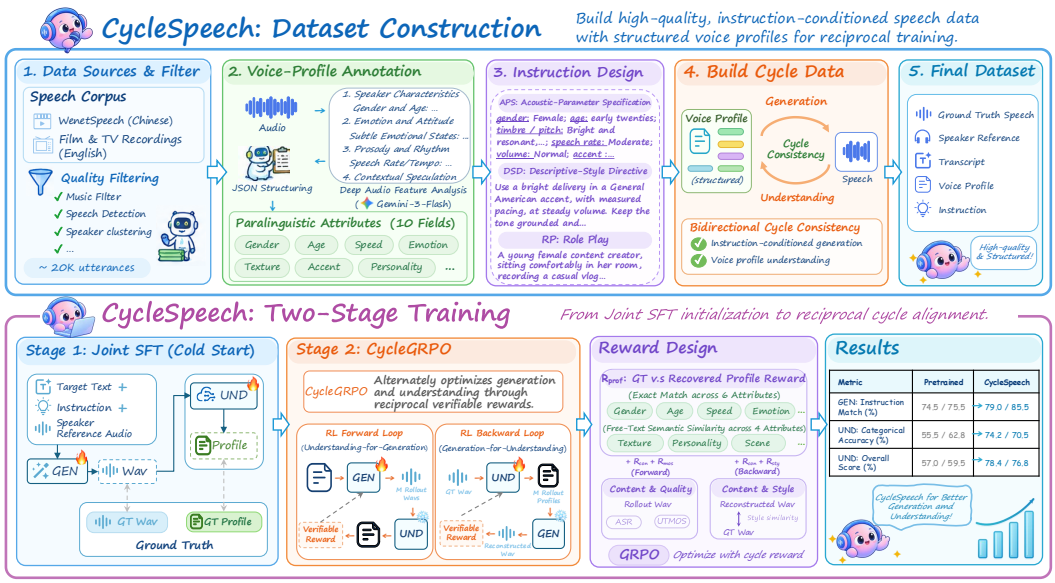

Figure 2: The Overview of CycleSpeech data construction and training framework. Joint SFT initializes instruction-conditioned generation (GEN) and speech-to-profile understanding (UND), after which CycleGRPO alternately optimizes generation and understanding from reciprocal cycles.

television recordings. After filtering music, non-speech, and multi-speaker segments, Gemini-3-flash Google DeepMind (2025) produces a qualitative report covering speaker characteristics, vocal texture, prosody, emotion, and inferred speaking context. DeepSeek-v3 Liu et al. (2024) for Chinese and Gemini-3-flash for English then normalize these reports into a shared JSON schema, shown in Table 1. Attribute-specific models refine demographics, emotion, accent, and volume. Because the source corpora lack speaker labels, we cluster WeSpeaker Wang et al. (2023a) embeddings within each session to obtain pseudo speaker IDs. A separate utterance with the same pseudo ID is selected as the speaker reference. This pipeline yields 20,046 bilingual examples; model-specific annotation and clustering settings are provided in the Appendix B.

Annotation consistency. Three annotators assessed 1,162 bilingual utterances (Table 2). Categorical attributes are evaluated against human labels, whereas descriptive attributes are rated for label–audio consistency on a 1–5 scale, with 5 indicating the closest match.

Table 1: Example of a voice profile. The schema captures speaker, prosodic, affective, and contextual attributes.

```
{
  "gender": "M",
  "age": "young",
  "accent": "mandarin",
  "emotion": "Neutral",
  "emotion_detail": ["informative", "quiet pride"],
  "personality": "Confident, resolute and...",
  "scene": "A leader speaking at an assembly...",
  "speed": "fast",
  "volume": "normal",
  "texture": "Bright, penetrating timbre..."
}
```

Table 2: Attribute-level consistency between human and model annotations. Categorical attributes are reported as accuracy (%), and descriptive attributes on a 1–5 scale.

Attribute	Consistency ↑	Attribute	Consistency ↑
Gender	96.7	Age	95.7
Accent	91.6	Emotion	70.8
Speed	86.7	Volume	72.5
Emo. Detail	4.92	Personality	4.95
Texture	4.87	Scene	4.79

3.3 Joint Supervised Fine-Tuning

Joint SFT establishes instruction-conditioned generation and speech-to-profile recovery, providing a supervised initialization for reciprocal reinforcement learning, as illustrated in Figure 2. With teacher forcing, GEN predicts A^* from (I, T, R), preserving the inference-time instruction interface. At each joint step, GEN samples $\widehat{A} \sim p_{\theta_G}(A \mid I, T, R; \phi)$, which the frozen token-to-wave decoder converts into $\widehat{x}$. UND recovers the canonical JSON sequence $Z(P^*)$ from this online-generated

waveform. Both branches use length-normalized autoregressive token negative log-likelihoods:

$$\mathcal{L}_{\mathrm{Gen}} = -\frac{1}{|A^*|}\sum_{t=1}^{|A^*|}\log p_{\theta_G}\left(a_t^* \mid I, T, R, a_{<t}^*; \phi\right),$$
$$\mathcal{L}_{\mathrm{Und}} = -\frac{1}{|Z(P^*)|}\sum_{m=1}^{|Z(P^*)|}\log p_{\theta_U}\left(z_m^* \mid \mathrm{sg}[\widehat{x}], z_{<m}^*; \phi\right). \tag{1}$$

After the GEN-only warm-up, the joint objective is:

$$\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{Gen}} + \lambda_{\mathrm{und}}\mathcal{L}_{\mathrm{Und}}, \qquad \lambda_{\mathrm{und}} = 1. \tag{2}$$

where sg stops gradients across the waveform interface, so $\mathcal{L}_{\mathrm{Gen}}$ and $\mathcal{L}_{\mathrm{Und}}$ update only θ_G and θ_U, respectively. Because UND is trained on the current generator's output distribution, unreliable early rollouts would provide an unstable acoustic input. We therefore optimize GEN alone for the first 2,000 updates, allowing it to produce usable instruction-conditioned speech before introducing online UND supervision. Both adapters are then jointly trained until step 10,000, exposing UND to the generator's evolving outputs.

3.4 CycleGRPO Reciprocal Alignment

CycleGRPO links two checks: whether generated speech realizes target attributes, and whether inferred profiles preserve information that can guide speaking-style reconstruction. Both adapters are initialized from Joint SFT before alternating optimization begins. During each update, the active adapter samples $M = 6$ candidates while the other supplies inference-only cycle feedback.

3.4.1 Forward Loop: Synthesizer Alignment

Naturalness alone does not establish that generated speech realizes the requested vocal attributes. The forward loop assesses attribute mismatches by recovering a profile from the waveform and comparing it with P^*. GEN samples $M = 6$ speech candidates from the instruction I, transcript T, and speaker reference R, without conditioning on P^*. The current UND adapter then deterministically recovers a voice profile from each waveform:

$$\widehat{A}_i \sim \pi_{\theta_G}(\cdot \mid I, T, R), \qquad \widehat{x}_i = \mathrm{Dec}_{\mathrm{token2wav}}(\widehat{A}_i; R),$$
$$\widehat{P}_i = \mathrm{UND}_{\theta_U}(\widehat{x}_i), \qquad i = 1, \ldots, M. \tag{3}$$

Profile consistency supplies the attribute-control signal, while content and quality terms penalize degradation in intelligibility and acoustic quality. The resulting GEN reward is a weighted combination of these complementary signals:

$$R_i^G = 0.70R_{\mathrm{prof}}(\widehat{P}_i, P^*) + 0.15\widetilde{R}_{\mathrm{con}}(\widehat{x}_i, T) + 0.15R_{\mathrm{mos}}(\widehat{x}_i). \tag{4}$$

Profile consistency. The canonical profile contains six categorical fields (gender, age, accent, emotion, speed, and volume) and four free-text fields (emotion detail, personality, scene, and vocal texture). List-valued descriptions are matched before aggregation:

$$R_{\mathrm{prof}} = \frac{1}{10}\left(\sum_{k\in\mathcal{K}_{\mathrm{cat}}} s_k + \sum_{k\in\mathcal{K}_{\mathrm{sem}}} s_k\right). \tag{5}$$

Here, $\mathcal{K}_{\mathrm{cat}}$ and $\mathcal{K}_{\mathrm{sem}}$ index the categorical and free-text fields, respectively. For field k, s_k is an exact-match indicator for categorical attributes or BGE-M3 Chen et al. (2024) cosine similarity for free-text attributes. The factor $1/10$ averages the ten paralinguistic fields equally.

Content and quality guards. English and Chinese candidates are transcribed with Whisper-large-v3 Radford et al. (2023) and FireRedASR2S Xu et al. (2026), respectively. Let ER denote WER or CER. The content reward saturates after high recognition accuracy:

$$R_{\mathrm{con}}^{\mathrm{raw}} = \mathrm{clip}(1 - \mathrm{ER}, 0, 1),$$
$$\widetilde{R}_{\mathrm{con}} = \min(R_{\mathrm{con}}^{\mathrm{raw}}/0.95, 1). \tag{6}$$

This preserves a penalty for genuine content degradation without over-optimizing small ASR differences. R_{mos} is the normalized UTMOS Saeki et al. (2022) score and discourages acoustically degraded solutions.

3.4.2 BACKWARD LOOP: COMPREHENSION ALIGNMENT

Because profile matching alone does not establish controllability, the backward loop tests whether inferred attributes support speaking-style reconstruction. UND samples $M = 6$ profiles from the gold waveform x^* and scores them against the fixed target P^* to anchor training. A deterministic bridge $\mathcal{B}$ converts each predicted profile and transcript T into a natural-language GEN input $\widetilde{I}_i$:

$$\begin{aligned} \widehat{P}_i &\sim \pi_{\theta_U}(\cdot \mid x^*), \qquad \widetilde{I}_i = \mathcal{B}(\widehat{P}_i, T), \\ \widetilde{A}_i &\sim \pi_{\theta_G}(\cdot \mid \widetilde{I}_i, R), \qquad \widetilde{x}_i = \mathrm{Dec}_{\mathrm{token2wav}}(\widetilde{A}_i; R). \end{aligned} \tag{7}$$

GEN receives the predicted profile through $\widetilde{I}_i$. The UND reward combines direct profile supervision with content and style feedback from the reconstructed speech:

$$R_i^U = 0.60 R_{\mathrm{prof}}(\widehat{P}_i, P^*) + 0.20 \widetilde{R}_{\mathrm{con}}(\widetilde{x}_i, T) + 0.20 R_{\mathrm{sty}}(\widetilde{x}_i, x^*). \tag{8}$$

The style term measures speaking-style similarity between the reconstructed and source speech. The content term guards transcript fidelity during reconstruction, with T supplied independently of the inferred profile.

Style reconstruction. We use a frozen ParaMETA encoder Lou et al. (2026) to measure the paralinguistic speaking-style similarity between the GEN-reconstructed waveform $\widetilde{x}_i$ and the ground-truth waveform x^*:

$$R_{\mathrm{sty}} = \mathrm{Sim}_{\mathrm{ParaMETA}}(\widetilde{x}_i, x^*). \tag{9}$$

where $\mathrm{Sim}_{\mathrm{ParaMETA}}$ denotes similarity between the ParaMETA speaking-style representations of the two waveforms. This reward is used only in the backward loop. During these updates, UND is the active policy and GEN remains inference-only, so the discrete profile and waveform interfaces prevent gradients from crossing between adapters. Further reward normalization details and bridge templates are provided in the Appendix.

3.4.3 ALTERNATING OPTIMIZATION

We construct the RL training pool using repeated Joint-SFT rollouts, discarding invalid, nearly solved, and reward-degenerate groups. The remaining prompts are stratified by difficulty, using mean reward as a proxy, and within-group reward variance. This yields 9,962 examples, comprising 4,976 GEN and 4,986 UND prompts.

On this pool, we alternate five GEN updates with five UND updates using the direction-specific rewards defined above. For each direction, a frozen copy of the corresponding Joint-SFT adapter serves as the reference policy for KL regularization.

Using R_i^G for GEN or R_i^U for UND, we compute a group-normalized sequence advantage A_i for each rollout. We optimize the active adapter using the same clipped GRPO objective:

$$\mathcal{L}_{\mathrm{GRPO}} = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{T_i} \sum_{t=1}^{T_i} \big[-\min\big(r_{i,t} A_i, \mathrm{clip}(r_{i,t}, 1-\epsilon, 1+\epsilon) A_i\big) + \beta \widehat{D}_{i,t} \big]. \tag{10}$$

where $r_{i,t}$ is the policy ratio and $\widehat{D}_{i,t}$ is the sampled-action KL penalty to the frozen Joint-SFT policy. The same A_i is applied to all T_i generated tokens in rollout i. We use clipping threshold $\epsilon = 0.2$ and KL coefficient $\beta = 0.04$ for both directions. Further implementation details are provided in the Appendix.

4 EXPERIMENT

4.1 EVALUATION SETUP AND METRICS

Test speech and instructions are held out from the curated dataset. The GEN evaluation set contains 1,077 prompts, balanced across APS, DSD, and RP within each language. The UND evaluation set comprises 2,127 source utterances: 1,144 Chinese and 983 English. We compare instruction-controlled generation and paralinguistic understanding baselines (Tables 3 and 4).

Automatic generation metrics. Content accuracy is measured by Chinese CER using FireRedASR2S Xu et al. (2026) and English WER using Whisper-large-v3 Radford et al. (2023). Speaker verification (Spk. Ver.) reports the percentage of generated/reference pairs with CAM++ Wang et al. (2023b) speaker-embedding cosine similarity of at least 0.33. Style Acc. measures exact-match accuracy over task-specific paralinguistic attributes including gender, age, accent, emotion, speed, and volume. Each applicable field receives a binary (0/1) exact-match score, averaged within each utterance. Style SIM measures cosine similarity between temporally pooled FACodec Ju et al. (2024) prosody representations of generated and ground-truth speech. Instruction Match reports the percentage of positive instruction–speech consistency judgments from Gemini-2.5-pro, using the evaluation prompt provided by InstructTTSEval Huang et al. (2025).

Speech understanding metrics. PR and SA measure profile parseability and field-level schema adherence. Cat.Avg evaluates gender, age, accent, emotion, speed, and volume, while Desc.Sim measures semantic similarity over personality, scene, voice texture, and emotion detail. TF measures transcript fidelity, and OS summarizes categorical, descriptive, and transcription quality while accounting for format adherence. Full definitions are provided in Appendix D.2.

5 Results and Analysis

Table 3: Instruction-controlled TTS results. WER/CER and Style Acc. use the shared evaluation protocol in Section 4.1; brackets give signed 95% confidence-bound offsets. Best and second-best results are **bolded** and underlined.

Method / Model	CER / WER(%) ↓	Spk. Ver.(%) ↑	Style Acc.(%) ↑	Style SIM↑	Instruction Match(%) ↑
Part I: Chinese Evaluation Benchmark					
GT Wav (upper reference)	4.44 [-0.99, +1.09]	99.00 [-0.83, +0.67]	62.56 [-2.59, +2.60]	1.00 [0.00, 0.00]	81.50 [-3.17, +3.00]
VoxInstruct Zhou et al. (2024)	10.90 [-1.61, +1.73]	80.17 [-3.17, +3.17]	58.87 [-2.37, +2.31]	0.87 [-0.01, +0.01]	73.00 [-3.50, +3.50]
CosyVoice3 Du et al. (2025)	2.76 [-0.87, +1.17]	**99.50** [-0.67, +0.50]	56.28 [-2.23, +2.21]	0.92 [-0.01, +0.01]	**82.17** [-3.17, +3.00]
OV-InstructTTS Ren et al. (2026)	3.64 [-0.76, +0.82]	99.00 [-0.83, +0.67]	58.63 [-2.21, +2.23]	0.90 [-0.01, +0.01]	79.67 [-3.33, +3.17]
MiMo-Audio-Instruct Zhang et al. (2025a)	4.42 [-0.77, +0.82]	81.83 [-3.00, +3.00]	58.76 [-2.13, +2.11]	0.92 [-0.01, +0.01]	80.33 [-3.17, +3.17]
Comparison Group A: Optimization Against Base Architecture					
Step-Audio-2-mini Wu et al. (2025)	**1.87** [-0.53, +0.59]	99.33 [-0.67, +0.50]	58.07 [-2.23, +2.23]	0.93 [-0.01, +0.01]	74.50 [-3.50, +3.50]
CycleSpeech (Phase 1: Joint SFT)	2.63 [-0.63, +0.71]	98.83 [-0.83, +0.83]	60.20 [-2.30, +2.31]	0.93 [-0.01, +0.01]	77.50 [-3.33, +3.33]
CycleSpeech (Phase 2: CycleGRPO)	2.33 [-0.56, +0.65]	**99.50** [-0.67, +0.50]	**60.42** [-2.30, +2.29]	**0.94** [-0.01, +0.01]	79.00 [-3.33, +3.33]
Part II: English Evaluation Benchmark					
GT Wav (upper reference)	6.47 [-1.73, +2.00]	100.00 [0.00, 0.00]	67.91 [-2.93, +2.87]	1.00 [0.00, 0.00]	82.39 [-3.56, +3.35]
VoxInstruct Zhou et al. (2024)	12.97 [-2.30, +2.52]	82.81 [-3.56, +3.35]	56.75 [-2.56, +2.66]	0.89 [-0.01, +0.01]	74.42 [-3.56, +3.35]
CosyVoice3 Du et al. (2025)	3.42 [-0.95, +1.07]	**99.79** [-0.42, +0.21]	58.15 [-2.74, +2.69]	0.93 [-0.01, 0.00]	80.08 [-3.56, +3.56]
OV-InstructTTS Ren et al. (2026)	11.02 [-5.63, +10.12]	93.29 [-2.31, +2.10]	55.44 [-2.50, +2.53]	0.89 [-0.02, +0.01]	78.41 [-3.56, +3.56]
MiMo-Audio-Instruct Zhang et al. (2025a)	7.00 [-1.28, +1.42]	72.54 [-4.19, +3.98]	56.37 [-2.56, +2.54]	0.91 [-0.01, +0.01]	77.57 [-3.77, +3.77]
Comparison Group A: Optimization Against Base Architecture					
Step-Audio-2-mini Wu et al. (2025)	**2.56** [-0.82, +0.95]	96.86 [-1.68, +1.46]	58.18 [-2.74, +2.72]	**0.94** [-0.01, +0.01]	75.47 [-3.77, +3.77]
CycleSpeech (Phase 1: Joint SFT)	2.74 [-0.80, +0.96]	98.95 [-1.05, +0.84]	60.35 [-2.39, +2.31]	0.92 [-0.01, +0.01]	81.97 [-3.56, +3.35]
CycleSpeech (Phase 2: CycleGRPO)	3.08 [-0.92, +1.07]	98.53 [-1.26, +1.05]	**60.78** [-2.41, +2.44]	0.92 [-0.01, +0.01]	**85.53** [-3.14, +3.14]

5.1 Main Quantitative Results

Speech Generation Table 3 shows complementary gains from the two training stages. Joint SFT improves Style Acc. over the backbone. CycleGRPO further strengthens instruction–speech consistency, increasing Instruction Match by 1.50/3.56 points on Chinese/English over Joint SFT. CycleSpeech achieves the highest Style Acc. in both languages and the highest English Instruction Match among the evaluated generation models, while maintaining strong speaker verification. Relative to Joint SFT, CycleGRPO reduces Chinese CER, with a slight increase in English WER.

Speech Understanding and Profile Recovery Joint SFT substantially improves profile recovery over Step-Audio-2-mini, raising Chinese/English OS by 20.98/17.09 points (Table 4). CycleGRPO further strengthens categorical and descriptive profile recovery, achieving the highest Cat.Avg, Desc.Sim, and OS in both languages while maintaining TF. Together with the generation results, these findings demonstrate stronger instruction control and speech-to-profile understanding.

5.2 Out-of-Domain Generalization

We evaluate cross-dataset generalization on the VccmDataset test set A Ji et al. (2024b). As shown in Table 5, *CycleSpeech (Phase 1)* consistently outperforms Step-Audio-2-mini across all paralin-

Table 4: Speech paralinguistic understanding results. **PR**: parse rate; **SA**: schema adherence; **Cat.Avg**: six-field categorical accuracy; **Desc.Sim**: four-field BGE-M3 cosine similarity; **TF**: transcription fidelity; **OS**: overall score. Best and second-best results are **bolded** and underlined.

Model	Structure (%) ↑		Attributes (%) ↑		Transcript (%) ↑	Overall (%) ↑
	PR	SA	Cat.Avg	Desc.Sim	TF	OS
Part I: Chinese Evaluation Benchmark						
Qwen2.5-Omni	**100.0**	93.1	58.52 [-1.03, +1.03]	45.37 [-0.45, +0.46]	85.9 [-0.82, +0.79]	58.76 [-0.56, +0.54]
MiMo-Audio-7B-Instruct	99.8	99.6	58.10 [-1.06, +1.02]	61.05 [-0.36, +0.35]	83.0 [-0.97, +0.93]	66.25 [-0.60, +0.58]
Kimi-Audio-7B-Instruct	94.4	90.9	45.37 [-1.43, +1.43]	53.54 [-0.86, +0.84]	80.5 [-1.49, +1.41]	56.26 [-1.06, +1.06]
Step-Audio-2-mini	99.0	91.9	55.46 [-1.03, +1.03]	52.12 [-0.60, +0.62]	77.8 [-1.62, +1.59]	57.00 [-0.76, +0.74]
CycleSpeech (Phase 1: Joint SFT)	**100.0**	**100.0**	73.56 [-1.08, +1.09]	71.90 [-0.42, +0.40]	**89.9** [-0.75, +0.72]	77.98 [-0.57, +0.56]
CycleSpeech (Phase 2: CycleGRPO)	**100.0**	**100.0**	**74.18** [-1.06, +1.08]	**72.30** [-0.42, +0.42]	**89.9** [-0.75, +0.73]	**78.36** [-0.56, +0.55]
Part II: English Evaluation Benchmark						
Qwen2.5-Omni	98.9	86.6	55.41 [-1.46, +1.44]	35.40 [-1.14, +1.10]	83.9 [-1.41, +1.33]	53.06 [-1.09, +1.05]
MiMo-Audio-7B-Instruct	99.8	97.3	65.26 [-1.19, +1.17]	58.95 [-0.60, +0.61]	87.0 [-1.00, +0.95]	68.35 [-0.68, +0.69]
Kimi-Audio-7B-Instruct	99.0	91.2	58.44 [-1.32, +1.34]	47.33 [-0.77, +0.79]	84.2 [-1.41, +1.38]	58.32 [-0.82, +0.83]
Step-Audio-2-mini	99.9	91.7	62.83 [-1.15, +1.15]	43.88 [-0.51, +0.52]	86.9 [-0.95, +0.93]	59.45 [-0.57, +0.58]
CycleSpeech (Phase 1: Joint SFT)	**100.0**	**100.0**	70.19 [-1.14, +1.14]	71.36 [-0.42, +0.41]	**90.3** [-0.91, +0.89]	76.54 [-0.59, +0.59]
CycleSpeech (Phase 2: CycleGRPO)	**100.0**	**100.0**	**70.50** [-1.14, +1.15]	**71.94** [-0.40, +0.40]	**90.3** [-0.91, +0.89]	**76.84** [-0.61, +0.60]

Table 5: Out-of-domain generation results across WER (%), Speaker Similarity (SIM), and Attribute Accuracies (%) for Pitch (PIT), Speed (SPD), Volume (VOL), and Emotion (EMO).

Model	WER↓	SIM↑	PIT↑	SPD↑	VOL↑	EMO↑
GT Codec	$3.47^{\pm 0.35}$	$.970^{\pm .002}$	$90.2^{\pm 1.5}$	$88.7^{\pm 1.6}$	$89.7^{\pm 1.5}$	$72.6^{\pm 6.3}$
Step-Audio-2-mini	$\mathbf{1.93}^{\pm 0.40}$	$\mathbf{.860}^{\pm .005}$	$69.7^{\pm 2.4}$	$65.6^{\pm 2.4}$	$63.2^{\pm 2.4}$	$33.7^{\pm 6.8}$
CycleSpeech (Phase1)	$2.66^{\pm 0.40}$	$.852^{\pm .005}$	$\mathbf{73.1}^{\pm 2.2}$	$67.1^{\pm 2.4}$	$65.6^{\pm 2.4}$	$37.4^{\pm 6.8}$
CycleSpeech (Phase2)	$2.50^{\pm 0.40}$	$.851^{\pm .006}$	$72.5^{\pm 2.3}$	$\mathbf{68.3}^{\pm 2.4}$	$\mathbf{66.0}^{\pm 2.4}$	$\mathbf{40.0}^{\pm 7.1}$

Table 6: Subjective MOS for instruction/style faithfulness (FMOS), naturalness (NMOS), and speaker similarity (SMOS).

Model	FMOS ↑	NMOS ↑	SMOS ↑
CosyVoice3	$3.82^{\pm .24}$	$3.91^{\pm .21}$	$3.71^{\pm .19}$
Step-Audio-2-mini	$3.61^{\pm .20}$	$3.88^{\pm .17}$	$3.53^{\pm .18}$
CycleSpeech (phase1)	$3.99^{\pm .21}$	$4.07^{\pm .21}$	$3.90^{\pm .22}$
CycleSpeech (phase2)	$\mathbf{4.15}^{\pm .21}$	$\mathbf{4.25}^{\pm .17}$	$\mathbf{4.00}^{\pm .20}$

guistic attributes. Building upon this, *CycleSpeech (Phase 2)* achieves the highest speed, volume, and emotion accuracies among the evaluated synthesis models, accompanied by lower WER relative to Phase 1. Speaker similarity remains competitive across both phases.

5.3 Subjective Evaluation

Five listeners evaluated 50 speech samples per system. Speech samples were presented in randomized order with system identities concealed. In the merged evaluation, each system received 100 ratings in total, corresponding to two ratings per sample on average. The superscripts in Table 6 indicate the half-widths of the 95% confidence intervals.

CycleSpeech achieves the highest mean subjective ratings across instruction/style faithfulness, naturalness, and speaker similarity (Table 6). The gains over Joint SFT, including a 0.18-point increase in NMOS, indicate that improved perceived style adherence need not compromise naturalness or speaker similarity.

5.4 Efficacy of CycleGRPO Closed-Loop Alignment

Alternating Optimization Dynamics. Alt-1/5/10 share reward weights, sampling settings, and the training pool. All alternating variants are evaluated at global step 1,300, corresponding to 650 updates per adapter. Profile recovery varies little across intervals, while generation is more sensitive to the switching frequency (Table 7). Alt-5 combines the lowest Chinese CER and highest Chinese Style Acc. with joint-highest OS in both languages, supporting its use as the default interval.

Impact of Reciprocal Cycle Rewards. The *w/o Cycle* variant retains alternating GRPO updates for both adapters but disables both feedback loops. Generated speech is not verified by UND, and predicted profiles are not used by GEN for reconstruction. GEN uses equally weighted ASR-based content and UTMOS quality rewards; UND uses only ground-truth profile matching. *w/ Frozen FB* retains both cycles, alternating updates of both adapters, and CycleGRPO's training budget, but obtains feedback from frozen Joint-SFT counterparts. The *w/o UND Updates* variant updates only GEN with UND frozen at Joint-SFT; *w/o GEN Updates* reverses these roles. Each retains the active

Table 7: Effect of the alternating update interval in CycleGRPO. Alt-k switches the optimized branch every k updates.

	GEN		UND		
Variant	**WER/CER (%) ↓**	**Style Acc. ↑**	**Cat. Avg. ↑**	**Desc.Sim ↑**	**OS ↑**
Chinese Test Set					
CycleGRPO (Alt-1)	2.47 [-0.60, +0.67]	58.82 [-2.30, +2.28]	74.08 [-1.06, +1.08]	**72.37** [-0.41, +0.40]	**78.36** [-0.56, +0.55]
CycleGRPO (Alt-10)	2.53 [-0.60, +0.66]	59.69 [-2.23, +2.21]	74.14 [-1.08, +1.05]	72.29 [-0.42, +0.40]	78.31 [-0.56, +0.55]
CycleGRPO (Alt-5, Ours)	**2.33** [-0.56, +0.65]	**60.42** [-2.30, +2.29]	**74.18** [-1.06, +1.08]	72.30 [-0.42, +0.42]	**78.36** [-0.56, +0.55]
English Test Set					
CycleGRPO (Alt-1)	**2.85** [-0.86, +1.01]	**61.21** [-2.50, +2.42]	70.30 [-1.14, +1.14]	**71.95** [-0.41, +0.40]	76.74 [-0.59, +0.58]
CycleGRPO (Alt-10)	3.21 [-0.92, +1.07]	59.55 [-2.66, +2.60]	**70.65** [-1.14, +1.12]	71.79 [-0.41, +0.40]	**76.84** [-0.59, +0.59]
CycleGRPO (Alt-5, Ours)	3.08 [-0.92, +1.07]	60.78 [-2.41, +2.44]	70.50 [-1.14, +1.15]	71.94 [-0.40, +0.40]	**76.84** [-0.61, +0.60]

Table 8: Effects of reciprocal cycle feedback, evolving feedback, and bidirectional updates. w/o Cycle removes both feedback loops but retains alternating updates ($k = 5$). w/ Frozen FB retains both cycles and alternating updates, with feedback from frozen Joint-SFT counterparts. w/o UND/GEN Updates train only GEN/UND, respectively, with the opposite adapter frozen at Joint-SFT.

	GEN		UND		
Variant	**WER/CER (%) ↓**	**Style Acc. ↑**	**Cat. Avg. ↑**	**Desc.Sim ↑**	**OS ↑**
Chinese Test Set					
w/o Cycle	**2.29** [-0.56, +0.61]	58.92 [-2.26, +2.18]	74.01 [-1.09, +1.09]	**72.45** [-0.41, +0.40]	78.35 [-0.56, +0.55]
w/ Frozen FB	2.74 [-0.70, +0.79]	59.66 [-2.24, +2.24]	74.08 [-1.08, +1.08]	72.36 [-0.40, +0.40]	78.34 [-0.56, +0.55]
w/o UND Updates	2.46 [-0.58, +0.65]	59.69 [-2.21, +2.17]	73.56 [-1.08, +1.09]	71.90 [-0.42, +0.40]	77.98 [-0.57, +0.56]
w/o GEN Updates	2.63 [-0.63, +0.71]	60.20 [-2.30, +2.31]	73.97 [-1.06, +1.08]	72.26 [-0.41, +0.40]	78.25 [-0.56, +0.56]
CycleGRPO (Ours)	2.33 [-0.56, +0.65]	**60.42** [-2.30, +2.29]	**74.18** [-1.06, +1.08]	72.30 [-0.42, +0.42]	**78.36** [-0.56, +0.55]
English Test Set					
w/o Cycle	2.89 [-0.93, +1.11]	58.96 [-2.53, +2.42]	70.19 [-1.14, +1.14]	71.91 [-0.41, +0.40]	76.69 [-0.60, +0.59]
w/ Frozen FB	2.91 [-0.89, +1.06]	60.68 [-2.55, +2.46]	70.26 [-1.15, +1.12]	71.92 [-0.43, +0.42]	76.72 [-0.61, +0.60]
w/o UND Updates	2.91 [-0.87, +1.03]	59.53 [-2.46, +2.46]	70.19 [-1.14, +1.14]	71.36 [-0.42, +0.41]	76.54 [-0.59, +0.59]
w/o GEN Updates	**2.74** [-0.80, +0.96]	60.35 [-2.39, +2.31]	**70.50** [-1.10, +1.12]	71.88 [-0.41, +0.41]	76.81 [-0.59, +0.58]
CycleGRPO (Ours)	3.08 [-0.92, +1.07]	**60.78** [-2.41, +2.44]	**70.50** [-1.14, +1.15]	**71.94** [-0.40, +0.40]	**76.84** [-0.61, +0.60]

adapter's full reward objective, weights, and KL regularization. Each single-direction control is evaluated after 650 updates, matching the corresponding adapter's update count in CycleGRPO.

CycleSpeech achieves the highest Style Acc. in both languages and the highest Cat.Avg, tying the English best (Table 8). Against *w/o Cycle*, Style Acc. improves by 1.50/1.82 points on Chinese/English, supporting the contribution of cycle feedback to attribute control. Both single-direction controls remain below CycleSpeech in Style Acc.; w/o GEN Updates ties its English Cat.Avg. Comparison with *w/ Frozen FB* isolates the value of evolving feedback, Frozen feedback improves Style Acc. over *w/o Cycle*, while evolving both counterparts further improves Style Acc. and Cat.Avg.

6 Conclusion

We presented CycleSpeech, which unifies instruction-controlled speech synthesis and paralinguistic understanding through a shared, recoverable voice profile. The forward cycle verifies intended attributes in generated speech, while the backward cycle tests whether inferred profiles support speaking-style reconstruction. Building on Joint SFT, CycleGRPO converts these checks into reciprocal training feedback without human preference annotations or a separately trained preference reward model. Our aligned bilingual corpus supports both understanding supervision and generation verification. Bilingual evaluations and controlled ablations show improved instruction following and categorical profile recovery, alongside strong subjective naturalness and speaker similarity. Together, these findings support reciprocal verification through recoverable attributes as a practical approach to controllable speech generation.

AI USE STATEMENT

We used Gemini-3-flash and DeepSeek-v3 for speech annotation and structured voice-profile construction, as described in the data construction pipeline. Gemini-2.5-pro was used to evaluate instruction–speech consistency under the reported evaluation protocol. Generative AI tools also provided feedback on experimental design and assisted with result interpretation, literature search, manuscript drafting and editing, and LaTeX formatting. Annotation quality was assessed through the human consistency study reported in the paper. We reviewed all AI-assisted work and checked numerical statements and interpretations against the experimental results. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

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. *arXiv preprint arXiv:2601.04656*, 2026.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. Audio-aware large language models as judges for speaking styles. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 467–480. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.25. URL https://aclanthology.org/2025.findings-emnlp.25/.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. Scaling rich style-prompted text-to-speech datasets, 2025. URL https://arxiv.org/abs/2503.04713.

Anuj Diwan, Eunsol Choi, and David Harwath. ParaSpeechCLAP: A dual-encoder speech-text model for rich stylistic language-audio pretraining. *arXiv preprint arXiv:2603.28737*, 2026. URL https://arxiv.org/abs/2603.28737.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Google DeepMind. Gemini 3 Flash: Model card. https://deepmind.google/models/model-cards/gemini-3-flash/, December 2025. Accessed: 2026-07-29.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. PromptTTS: Controllable text-to-speech with text descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096285. URL https://arxiv.org/abs/2211.12171.

Takaaki Hori, Ramon Astudillo, Tomoki Hayashi, Yu Zhang, Shinji Watanabe, and Jonathan Le Roux. Cycle-consistency training for end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6271–6275. IEEE, 2019. doi: 10.1109/ICASSP.2019.8683307. URL https://arxiv.org/abs/1811.01690.

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. *arXiv preprint arXiv:2506.16381*, 2025. URL https://arxiv.org/abs/2506.16381.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. TextrolSpeech: A text style control speech corpus with codec language text-to-speech models. *arXiv preprint arXiv:2308.14430*, 2024a. URL https://arxiv.org/abs/2308.14430.

Shengpeng Ji, Jialong Zuo, Wen Wang, Minghui Fang, Siqi Zheng, Qian Chen, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, and Zhou Zhao. ControlSpeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec. *arXiv preprint arXiv:2406.01205*, 2024b. URL https://arxiv.org/abs/2406.01205.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. SpeechCraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, 2024. doi: 10.1145/3664647.3681674. URL https://arxiv.org/abs/2408.13608.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22605–22623. PMLR, 2024. URL https://proceedings.mlr.press/v235/ju24b.html.

Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana. Libritts-p: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning. *arXiv preprint arXiv:2406.07969*, 2024.

Shijia Liao, Yuxuan Wang, Songting Liu, Yifan Cheng, Ruoyi Zhang, Tianyu Li, Shidong Li, Yisheng Zheng, Xingwei Liu, Qingzheng Wang, Zhizhuo Zhou, Jiahua Liu, Xin Chen, and Dawei Han. Fish audio s2 technical report. *arXiv preprint arXiv:2603.08823*, 2026. URL https://arxiv.org/abs/2603.08823.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Haowei Lou, Hye young Paik, Wen Hu, and Lina Yao. ParaMETA: Towards learning disentangled paralinguistic speaking styles representations from speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32311–32319. AAAI Press, 2026. doi: 10.1609/aaai.v40i38.40505. URL https://doi.org/10.1609/aaai.v40i38.40505.

Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations, 2024. URL https://arxiv.org/abs/2402.01912.

Naoki Makishima, Satoshi Suzuki, Atsushi Ando, and Ryo Masumura. Speaker consistency loss and step-wise optimization for semi-supervised joint training of TTS and ASR using unpaired text data. In *Interspeech 2022*, pp. 526–530, 2022. doi: 10.21437/Interspeech.2022-304.

Junyi Peng, Oldřich Plchot, Xiao Song, Dading Chong, Lichun Fan, Hang Su, Themos Stafylakis, Junjie Li, Kong Aik Lee, Shuai Wang, Jian Luan, and Jan Černocký. SpeakerCard-1M: An evidence-grounded corpus for in-the-wild speaker verification. *arXiv preprint arXiv:2606.03283*, 2026. URL https://arxiv.org/abs/2606.03283.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Yong Ren, Jiangyan Yi, Jianhua Tao, Haiyang Sun, Zhengqi Wen, Hao Gu, Le Xu, and Ye Bai. OV-InstructTTS: Towards open-vocabulary instruct text-to-speech. *arXiv preprint arXiv:2601.01459*, 2026. URL https://arxiv.org/abs/2601.01459.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech 2022*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Listening while speaking: Speech chain by deep learning. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 301–308, 2017. doi: 10.1109/ASRU.2017.8268950. URL https://arxiv.org/abs/1707.04879.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Helin Wang, Jiarui Hai, Dading Chong, Karan Thakkar, Tiantian Feng, Dongchao Yang, Junhyeok Lee, Laureano Moro Velazquez, Jesus Villalba, Zengyi Qin, Shrikanth Narayanan, Mounya Elhiali, and Najim Dehak. Capspeech: Enabling downstream applications in style-captioned text-to-speech, 2025. URL https://arxiv.org/abs/2506.02863.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*, 2023b.

Yue Wang, Ruotian Ma, Xingyu Chen, Zhengliang Shi, Morunliu Yang, Wanshun Chen, Huang Liu, Jiadi Yao, Xin He, Qu Yang, Qingxuan Jiang, Fanghua Ye, Juntao Li, Zhaopeng Tu, Xiaolong Li, Liefeng Bo, and Min Zhang. BatonVoice: An operationalist framework for enhancing controllable speech synthesis with linguistic intelligence from LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 46683–46697. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2165. URL https://aclanthology.org/2026.acl-long.2165/.

Matt Whitehill, Shuang Ma, Daniel McDuff, and Yale Song. Multi-reference neural TTS stylization with adversarial cycle consistency. In *Interspeech 2020*, pp. 4442–4446, 2020. doi: 10.21437/Interspeech.2020-2985.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Hao Nie, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, Yuxiang Zhang, Zidong Yang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-audio 2 technical report, 2025. URL https://arxiv.org/abs/2507.16632.

Kaituo Xu, Yan Jia, Kai Huang, Junjie Chen, Wenpeng Li, Kun Liu, Feng-Long Xie, Xu Tang, and Yao Hu. Fireredasr2s: A state-of-the-art industrial-grade all-in-one automatic speech recognition system. *arXiv preprint arXiv:2603.10420*, 2026.

Liumeng Xue, Shifeng Pan, Lei He, Lei Xie, and Frank K. Soong. Cycle consistent network for end-to-end style transfer TTS training. *Neural Networks*, 140:223–236, 2021. doi: 10.1016/j.neunet.2021.03.005.

Kazuki Yamauchi, Yusuke Ijima, and Yuki Saito. StyleCap: Automatic speaking-style captioning from speech based on speech and language self-supervised learning models. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11261–11265. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10445977. URL https://doi.org/10.1109/ICASSP48485.2024.10445977.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024. doi: 10.1109/TASLP.2024.3402088. URL https://arxiv.org/abs/2301.13662.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, et al. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10748–10757, 2025.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025a.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, and Mingbo Ma. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2502.07243.

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*, pp. 554–563, 2024.

Xinfa Zhu, Wenjie Tian, Xinsheng Wang, Lei He, Yujia Xiao, Xi Wang, Xu Tan, Sheng Zhao, and Lei Xie. UniStyle: Unified style modeling for speaking style captioning and stylistic speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. doi: 10.1145/3664647.3681465. URL https://doi.org/10.1145/3664647.3681465.

A Paradigm Comparison

Table S1 compares the functional roles of generation, understanding, shared targets, and reciprocal optimization across paradigms.

Table S1: Paradigm comparison across RL-aligned instruction-controlled generation and paralinguistic evaluation or understanding. *Shared Target* denotes a semantic target shared by generation and understanding; *Cycle Opt.* denotes reciprocal optimization between them.

Method	Gen.	Und.	Shared Target	Cycle Opt.	Reward
RL-aligned generation					
CosyVoice 3	✓	✗	✗	✗	ASR/SER scores
Fish Audio S2	✓	✗	✗	✗	ASR, quality, speaker
FlexiVoice	✓	✗	✗	✗	Emotion and timbre preference
Evaluation and understanding					
InstructTTSEval	✗	✓	✗	✗	External audio-LLM
ALLM Judges	✗	✓	✗	✗	Instruction and style rating
StyleCap	✗	✓	✗	✗	Free-form caption
ParaSpeechCLAP	✗	✓	✗	✗	Speech–caption similarity
Joint generation and understanding					
UniStyle	✓	✓	✓	✗	Free-form caption
CycleSpeech	✓	✓	✓	✓	**Structured profile**

B Training Data Construction

A paired training instance is $\mathcal{D}_i = (I_i, T_i, R_i, x_i^*, A_i^*, P_i^*)$, containing an instruction, transcript, speaker reference, target waveform, target speech tokens, and reference voice profile, respectively. The shared backbone has frozen parameters ϕ; θ_G and θ_U parameterize the GEN and UND adapters. Figure S1 summarizes the annotation pipeline.

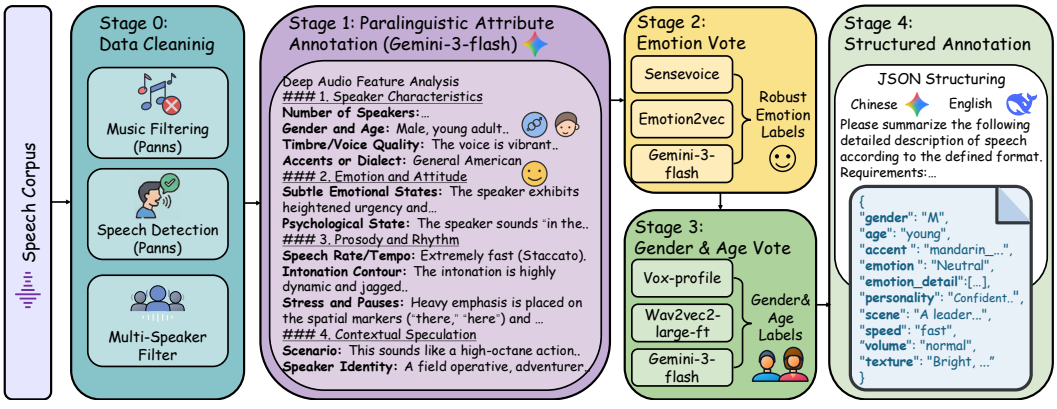

Figure S1: Data annotation pipeline.

B.1 Data Sources, Filtering, and Composition

Chinese speech is drawn from WenetSpeech, and English speech from an internal collection of film and television recordings. Full-length English recordings are first segmented using subtitle timestamps, and the corresponding subtitle text serves as the transcript.

Audio is converted to 16-kHz mono. We use PANNs Cnn14 to filter audio clips, retaining only those with a predicted speech probability of at least 0.5 and a predicted music probability below 0.5. We retain only single-speaker audio clips, excluding those containing multiple speakers.

The paired SFT set contains 20,046 source utterances. The GRPO pool contains 9,962 direction-specific records, comprising 4,976 GEN and 4,986 UND prompts. Table S2 summarizes their instruction counts and session-local pseudo speaker IDs. The observed duration ranges are 1.00–9.98 s for SFT and 1.00–10.152 s for GRPO.

Table S2: Training-set composition.

Stage	APS	DSD	RP	Total	Pseudo IDs
Joint SFT	6,047	8,266	5,733	20,046	17,043
CycleGRPO	3,791	3,571	2,600	9,962	7,996

B.2 Profile Annotation and Canonical Fields

Gemini-3-flash produces an auditory report covering speaker characteristics, affect, prosody, vocal texture, speaking behavior, and inferred communicative context. DeepSeek-v3 and Gemini-3-flash convert the Chinese and English reports, respectively, into structured voice profiles following a shared JSON schema.

Attribute refinement. Gender and age labels are refined by majority voting among predictions from Gemini-3-flash, WavLM-large-age-sex (Vox-Profile) and wav2vec2-large-robust-24-ft-age-gender. Emotion predictions from Gemini-3-flash, SenseVoice, and Emotion2Vec are mapped to seven core categories plus an "Other" category. Majority voting combines these predictions, with ties resolved using the Gemini-3-flash annotation. FireRedLID supplies Chinese accent or dialect labels, whereas English accent labels are taken directly from Gemini-3-flash annotations. The canonical label sets are listed in Table S3.

Table S3: Canonical categorical label sets.

Field	Canonical values
Gender	`M`, `F`, `unk`
Age	`child`, `teen`, `young`, `middle`, `senior`, `unknown`
Speed	`slow`, `medium_slow`, `medium`, `medium_fast`, `fast`, `unknown`
Volume	`quiet`, `normal`, `loud`, `unknown`
Emotion	`Neutral`, `Happy`, `Sad`, `Angry`, `Surprised`, `Fearful`, `Disgusted`, `Other`
English accent	`english_general_american`, `english_british`, `english_australian`, `english_indian`, `english_mid_atlantic`, `english_other`, `unknown`
Chinese accent	`mandarin`; `regional_` followed by `northern`, `beijing`, `northeast`, `taiwan`, `cantonese`, `southwest`, `hongkong`, `shaanxi`, `hunan`, `henan`, `sichuan`, `shandong`, `wu`, `min`, `other`, or `shanghai`; `unknown`

Profile fields and annotation metadata. The profile reward scores six categorical fields (gender, age, accent, emotion, speed, and volume) and four descriptive fields (emotion detail, personality, scene, and texture). The `emotion_detail` field is an array of one to three strings; the other descriptive fields are free text. SA evaluates 11 normalized fields: these ten paralinguistic fields and the transcript field. TF evaluates the predicted transcript.

B.3 Speaker Annotation and Reference Construction

WeSpeaker uses CnCeleb- and VoxCeleb-trained ResNet34 models for Chinese and English, respectively. Utterances of at most 15 s use one utterance-level embedding. Longer recordings use sliding-window embeddings with spectral clustering. Embeddings are then clustered within their source session using cosine-similarity thresholds of 0.55 for Chinese and 0.65 for English.

For each SFT target, reference selection first searches for another utterance with the same pseudo ID in the selected dataset. If none exists, it searches the full speaker mapping. The first eligible candidate is selected, excluding the target itself. Targets without an independent same-speaker reference

are excluded from the paired SFT set. All 20,046 retained SFT pairs have a nonempty reference that differs from the target and shares its pseudo speaker ID.

B.4 Instruction Construction

Instructions are rendered deterministically from selected profile fields, without an additional LLM call. Table S4 shows representative English templates; Chinese instructions use corresponding language-specific renderings.

For DSD, one of 20 language-specific templates is deterministically selected from the normalized profile. APS uses fixed slots, whereas RP directly adopts the free-form scene description. Instantiated bilingual examples appear in Table S5.

Table S4: Profile-to-Instruction templates. Braces denote values from the structured voice profile.

Type	Construction template
APS	gender: {gender}; age: {age}; timbre / pitch: {texture}; speech rate: {speed}; volume: {volume}; accent annotation: {accent}.
DSD	Speak at a {speed} pace, with {texture} timbre, at {volume} volume. Convey {emotion_detail}, in a {personality} manner.
RP	{scene}

Table S5: Examples of Chinese and English APS, DSD, and RP instructions for acoustic attributes, speaking styles, and communicative scenarios, respectively.

Source	Type	Instruction
Chinese	**APS**	女性，青壮年（20-30岁）；音色与音高偏高、清脆、明亮、金属质感；语速偏快，音量中等偏强；标准普通话。
	DSD	以中等音高为主，音色上体现柔和、纤细、湿润，音量极低，展现出这段语音带有明显的焦虑和紧迫感与秘密感交织的情感层次，性格层面显得紧张、脆弱、警示者，音频中，说话人语速较快，带有宿命般的强调。
	RP	在古装剧或宫廷题材场景中，一位青年女性角色在私密室内与心腹密探对话，听到的惊人消息后，她眉头微蹙、眼神凝滞，缓慢而沉重地复述这一事实，声音中充满震惊与疑虑，试图确认核实。
English	**APS**	gender: Female; age: early to mid-twenties; timbre / pitch: Bright and resonant, with a gentle terminal fall, Crystalline, smooth, slightly breathy, with a subtle hint of nasality; speech rate: Moderate and measured; volume: Normal conversational level; accent annotation: Standard General American accent.
	DSD	Use a bright delivery in a General American accent, in a mid register, with measured pacing, at steady volume. Keep the tone grounded and matter-of-fact, with a relaxed presence.
	RP	a young female content creator in her early twenties, sitting comfortably in her room, recording a casual vlog update for her followers.

C Additional Training and Reward Details

C.1 Joint SFT Sampling and Warm-Start Schedule

The token-level negative log-likelihoods $\mathcal{L}_{\mathrm{Gen}}$ and $\mathcal{L}_{\mathrm{Und}}$ are defined in Section 3.3 of the main paper. $\mathcal{L}_{\mathrm{Und}}$ denotes supervision on online-generated speech, with $\mathrm{sg}[\widehat{x}]$ as input. The frozen token-to-wave decoder converts a GEN rollout into a waveform:

$$\widehat{A} \sim p_{\theta_G}(A \mid I, T, R; \phi), \qquad \widehat{x} = \mathrm{Dec}_{\mathrm{token2wav}}(\widehat{A}; R). \tag{S1}$$

Rollouts use temperature 0.7, top-$p = 1.0$, top-$k = 50$, and repetition penalty 1.05. Their token budget depends on transcript length and is capped at 2,048 new tokens. UND uses the official Step-Audio-2 chat template. GEN first synthesizes $\widehat{x}$ without gradient tracking. UND is trained on the

detached waveform $\mathrm{sg}[\widehat{x}]$, so its loss updates only the UND adapter and does not backpropagate into GEN.

Let S_G and S_U denote GEN-only warm-up updates and optional UND-only warm-up updates on gold speech. At training step s, the schedule is

$$\mathcal{L}_{\mathrm{SFT}}^{(s)} = \begin{cases} \mathcal{L}_{\mathrm{Gen}}, & 0 \le s < S_G, \\ \lambda_{\mathrm{und}} \mathcal{L}_{\mathrm{Und}}^{\mathrm{gold}}, & S_G \le s < S_G + S_U, \\ \mathcal{L}_{\mathrm{Gen}} + \lambda_{\mathrm{und}} \mathcal{L}_{\mathrm{Und}}, & s \ge S_G + S_U. \end{cases} \tag{S2}$$

Here, $\mathcal{L}_{\mathrm{Und}}^{\mathrm{gold}}$ uses the ground-truth waveform x^*, rather than the detached GEN waveform $\mathrm{sg}[\widehat{x}]$, as input to UND. The selected setting is $S_G = 2{,}000$, $S_U = 0$, and $\lambda_{\mathrm{und}} = 1$.

Both adapters use LoRA rank 8 and dropout 0.05. GEN targets all linear layers with LoRA scaling parameter $\alpha = 32$; UND targets self-attention $q/k/v/o$ projections with $\alpha = 16$. AdamW uses learning rate 2×10^{-5}, weight decay 0.01, gradient clipping at 1.0, a 2,000-step linear learning-rate warm-up, and cosine decay. Learning-rate warm-up and the GEN-only warm-up are distinct schedules, although both last 2,000 steps. Training uses BF16 on four A800 GPUs with batch size 1 per worker and global batch size 4. Maximum sequence lengths are 6,144 for GEN and 2,048 for UND. The selected 10k-step Joint-SFT checkpoint initializes both CycleGRPO adapters from their respective weights.

C.2 CycleGRPO Optimization

During online training, we alternately update GEN and UND, switching adapters every $k = 5$ updates. Within each block, only the active adapter is optimized; the other provides cycle feedback using its latest weights without parameter updates. When providing feedback to GEN, UND decodes profiles deterministically. During its own update blocks, UND instead samples candidate profiles for GRPO. Separate Joint-SFT copies, π_G^{ref} and π_U^{ref}, remain frozen throughout training and are used only for KL regularization.

Each of four data-parallel workers samples $M = 6$ candidates for one prompt, giving 24 candidates per synchronized global step. For the active direction, R_i denotes R_i^G or R_i^U as defined in the main paper. Statistics are computed within each six-candidate prompt group, not across unrelated prompts:

$$\begin{aligned} \bar{R} &= \frac{1}{M}\sum_{i=1}^{M} R_i, \qquad \sigma_R = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(R_i - \bar{R})^2}, \\ \Delta_R &= \max_i R_i - \min_i R_i, \qquad A_i = \frac{R_i - \bar{R}}{\max(\sigma_R, 0.02)}. \end{aligned} \tag{S3}$$

A group is resampled without an optimizer update if $\sigma_R < 0.01$ or $\Delta_R < 0.02$. The standard deviation uses the population denominator M; the advantage denominator has a floor of 0.02. The sequence-level A_i is shared by all $T_i = |y_i|$ generated tokens in rollout i. Here, T_i is an output length, whereas T denotes the transcript.

The GRPO objective uses the following policy ratio and sampled-action KL estimator. All token probabilities condition on the direction-specific input and preceding generated tokens; these conditions are suppressed for brevity:

$$\begin{aligned} r_{i,t} &= \exp\left[\log \pi_\theta(y_{i,t}) - \log \pi_\theta^{\mathrm{old}}(y_{i,t})\right], \\ \delta_{i,t} &= \log \pi^{\mathrm{ref}}(y_{i,t}) - \log \pi_\theta(y_{i,t}), \qquad \widehat{D}_{i,t} = \exp(\delta_{i,t}) - \delta_{i,t} - 1. \end{aligned} \tag{S4}$$

The parameter θ is θ_G or θ_U for the active direction; $\pi_\theta^{\mathrm{old}}$ is the policy before the update and π^{ref} its frozen Joint-SFT reference. We use $\epsilon = 0.2$ and $\beta = 0.04$ for both directions.

GEN and UND use separate AdamW optimizers. Gradients are averaged across the four workers and clipped to norm 1 before updating the active adapter. CycleGRPO uses learning rate 5×10^{-7}, shuffled sampling, one update per accepted group, at most 32 resampling attempts.

C.3 Reward Implementation and Profile-to-Instruction Bridge

The profile and content reward formulas are given in Section 3.4 of the main paper. For a categorical field k, s_k is the exact-match indicator between the predicted and reference values. Descriptive fields use a frozen BGE-M3 encoder, with list-valued descriptions matched before aggregation. All ten paralinguistic fields have equal weight.

English and Chinese content rewards use Whisper-large-v3 and FireRedASR2S, respectively. WER/CER enter the reward as fractions. For raw UTMOS22-Strong prediction q, quality is normalized as:

$$R_{\text{mos}} = \text{clip}\left(\frac{q-1}{4}, 0, 1\right). \tag{S5}$$

The deterministic bridge $\mathcal{B}$ renders the predicted profile as a natural-language instruction and combines it with transcript T. Its template is selected by the instruction category; this category dependence is suppressed in the main-paper notation $\widetilde{I}_i = \mathcal{B}(\widehat{P}_i, T)$. GEN then uses $\widetilde{I}_i$ and the original speaker reference R to generate tokens $\widetilde{A}_i$, which are decoded into the reconstructed waveform $\widetilde{x}_i$. The backward-loop R_{sty} compares $\widetilde{x}_i$ with x^* using the frozen ParaMETA representation.

D Evaluation Protocols

D.1 Speech Generation Metrics

Content accuracy. We apply canonical text normalization to both ASR hypotheses and reference transcripts. Chinese CER and English WER are computed as the total edit distance divided by the total reference character or word count, respectively, and reported as percentages.

Style accuracy. Style Acc. evaluates gender, age, accent, speed, and volume for APS; emotion, speed, and volume for DSD; and gender, age, and emotion for RP. Each applicable field receives a binary exact-match score. Scores are averaged within each utterance, then within each instruction category, and finally equally across APS, DSD, and RP.

Attribute evaluators. For Style Acc., VoxProfile predicts gender and age, while VoxLect and FireRedLID predict English and Chinese accents, respectively. Emotion is assessed by a speech emotion recognition (SER) model built on the Qwen3-Omni Captioner audio encoder and trained on large-scale emotion data. Speaking rate is estimated using Voice-Taxonomy TEMP, and volume is measured by mean-frame root-mean-square (RMS) amplitude.

D.2 Speech Understanding Metrics

Parse rate and schema adherence. PR is the percentage of outputs parsed into profile objects. The format score F_i is the fraction of 11 normalized fields satisfying the schema: ten paralinguistic fields and the transcript, excluding language metadata. Unparseable outputs receive $F_i = 0$. Schema adherence is $\text{SA} = 100N^{-1}\sum_i F_i$, where N denotes the number of evaluated utterances.

Categorical and descriptive accuracy. C_i is the mean binary exact-match score over gender, age, accent, emotion, speed, and volume, multiplied by 100. D_i is 100 times the mean BGE-M3 Chen et al. (2024) cosine similarity over personality, scene, voice texture, and emotion detail. Cat.Avg and Desc.Sim average C_i and D_i, respectively, across utterances.

Transcript fidelity. TF averages $T_i = 100\ \text{clip}(1 - \text{CER}_i, 0, 1)$ across utterances.

Overall score. OS combines these per-utterance scores with format adherence:

$$\text{OS} = \frac{1}{N}\sum_{i=1}^{N} F_i\left(0.42C_i + 0.28D_i + 0.30T_i\right). \tag{S6}$$

D.3 Confidence Intervals

For in-domain metrics, 95% confidence intervals are estimated using 10,000 paired source-audio bootstrap resamples. The same source indices are sampled across systems, with all three GEN instruction views retained together for each sampled source.

D.4 Instruction Match Evaluation Prompt

Instruction Match uses Gemini-2.5-pro with the InstructTTSEval-based evaluation prompt. The prompt is reproduced below to preserve the evaluation criteria and requested output format.

Instruction Match Evaluation Prompt

You are an expert with rich knowledge of acoustics. Please describe an audio clip based on the following dimensions and judge whether the audio matches the provided description. Output **true** (matches) or **false** (does not match) on the consistency dimension, ignoring non-style factors (audio quality, naturalness, etc.).

Evaluation Dimensions

- **Gender:** Speech characteristics associated with different gender identities, including vocal fold differences and socialized speech patterns.
- **Pitch:** The perceived frequency of the sound, determining whether the voice sounds high or low. Typically, male voices have lower pitch and female voices have higher pitch. Relative pitch levels can be expressed based on gender, such as "female high pitch" or "male deep and steady pitch".
- **Speaking Rate:** The speed of speech, which often varies during a conversation. If the speaker demonstrates a specific rhythmic pattern, please specify it.
- **Volume:** The loudness or softness of speech, which may vary significantly. Examples include whispering, normal conversational volume, or shouting.
- **Age:** Infer the age group or life stage of the speaker (e.g., child, teenager, young adult, middle-aged, elderly) based on speech characteristics. If a specific stage is difficult to determine, simply indicate an approximate stage.
- **Accent:** A distinctive way of pronunciation reflecting geographical origin, socioeconomic background, or non-native speaker status. If the accent is sufficiently distinct, specify the dialect region as precisely as possible; otherwise, indicate a general region, such as American English, British English, or Mandarin.
- **Timbre & Texture:** The tonal quality and texture of the voice, including descriptions such as sweet, raspy, deep, bright, warm, nasal, soft, harsh, or delicate. These attributes reflect physiological characteristics (e.g., vocal cord structure) and stylistic nuances, which can be used to distinguish speakers or analyze emotional/expressive tendencies.
- **Emotion:** The emotion expressed while speaking, which may change during speech. For example, a person might start speaking calmly but gradually become frustrated, or switch from sadness to laughter within the same sentence.
- **Intonation / Tone:** The emotional or attitudinal nuance conveyed through vocal inflections, including pitch variation patterns, expressing subtle nuances such as sarcasm, formality, enthusiasm, or indifference.
- **Personality:** Infer the speaker's overall personality based on the above speech characteristics, such as extroverted/introverted, confident, decisive, or anxious. Describe only prominent and consistent personality traits present in the speech.
- **Role and Scenario Performance:** Whether the role state, relationship identity, and scenario atmosphere presented by the speech match the description, such as a caregiver's restraint and exhaustion, a sense of boundaries during an argument, or professionalism in a formal broadcast scenario. Evaluate only the performance style and state that can be perceived through sound, and do not judge information that cannot be verified by audio alone (e.g., specific location, physical appearance, actions, interpersonal relationships).

Judgment Criteria

Judgment	Definition
true (Matches)	The audio sample satisfies the main stylistic features of the description: • Primary style dimensions (e.g., gender, pitch, speaking rate, emotion) are consistent with the description • No obvious deviations or conflicts
false (Does not match)	The audio sample fails to satisfy the main stylistic features of the description: • At least one key stylistic feature obviously conflicts with the description • The overall auditory perception deviates from the described style

Precautions

- When there is an obvious and objective discrepancy in a certain dimension, such as speaker gender or age conflicting with the description, judge directly as false.
- There is a high probability that the description obviously conflicts with, differs in degree from, or completely mismatches the audio. Do not easily trust the provided description; hold your own independent understanding of the audio first.
- Pay special attention to speaker gender, as it is particularly prone to being opposite to the description.
- Pay extra attention when words indicating degree (such as "excited" or "intense") are mentioned in the description. The specified dimension in the audio may not be as intense as described (e.g., emotions are not excited enough, pitch is not high enough, or volume is not loud enough); in such cases, judge as false.
- Evaluate only **stylistic consistency**, ignoring non-style factors such as pronunciation accuracy or naturalness.
- Use the description as the sole basis without personal subjective bias.
- Features not mentioned in the description have no restrictions and should not affect the judgment.
- When the description focuses only on a specific dimension (e.g., emotion), judgment should focus primarily on that dimension.

Output Format Requirements:

Return a JSON object containing all fields below, replacing the placeholder strings with audio-based descriptions. Set `Consistency` to the Boolean `true` or `false` according to the judgment criteria; the value shown below is illustrative.

```
{
    "Gender": ...,
    "Pitch": ...,
    "Speaking Rate": ...,
    ...
    "Consistency": true
}
```

Description and Audio to be Evaluated:

<Insert description of audio style to be evaluated here>

D.5 Subjective Evaluation Protocol

Five listeners participated in the subjective evaluation. For each system, we evaluated 50 utterances, evenly split between Chinese and English (25 each), and each utterance received ratings from at least two listeners. Samples were anonymized and presented in randomized order. Faithfulness (FMOS), naturalness (NMOS), and speaker similarity (SMOS) were each rated on a five-point scale. Figure S2 shows the evaluation interface.

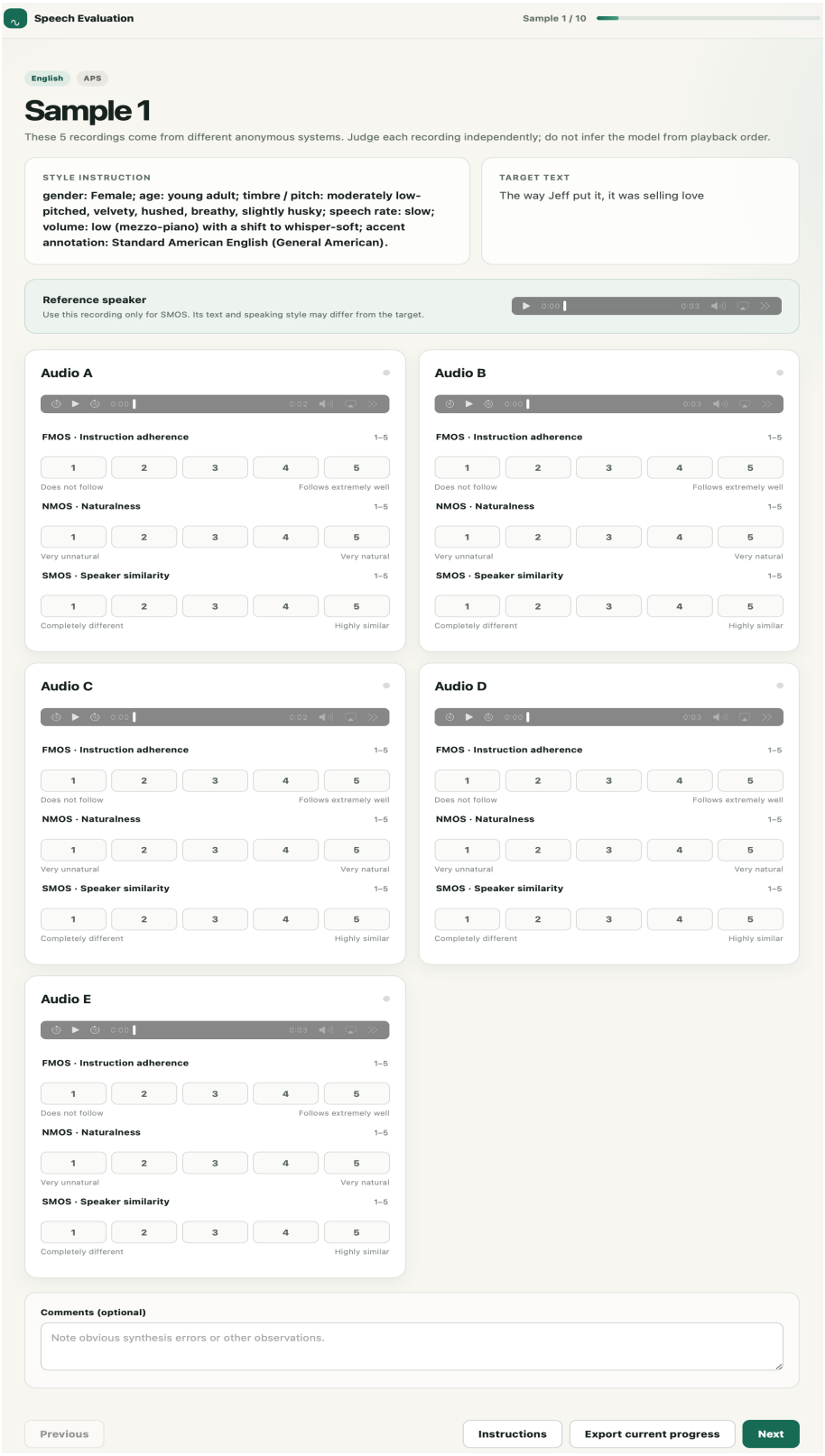

Figure S2: Listening-test interface for instruction faithfulness, naturalness, and speaker similarity.

E ADDITIONAL ABLATIONS

E.1 STAGE-I SFT ABLATION

Experimental setup. We compare Independent SFT (A1), which trains GEN and UND separately, with three Joint-SFT schedules: `GEN0/UND0`, `GEN2k/UND0`, and `GEN2k/UND1k`. The notation `GENx/UNDy` indicates the branch-specific warm-up steps used before joint training. Across the four configurations, we evaluate checkpoints between 1k and 10k steps on a fixed set, balanced between Chinese and English. For each case, both GEN and UND produce six candidates under the same reference annotations and scoring protocol.

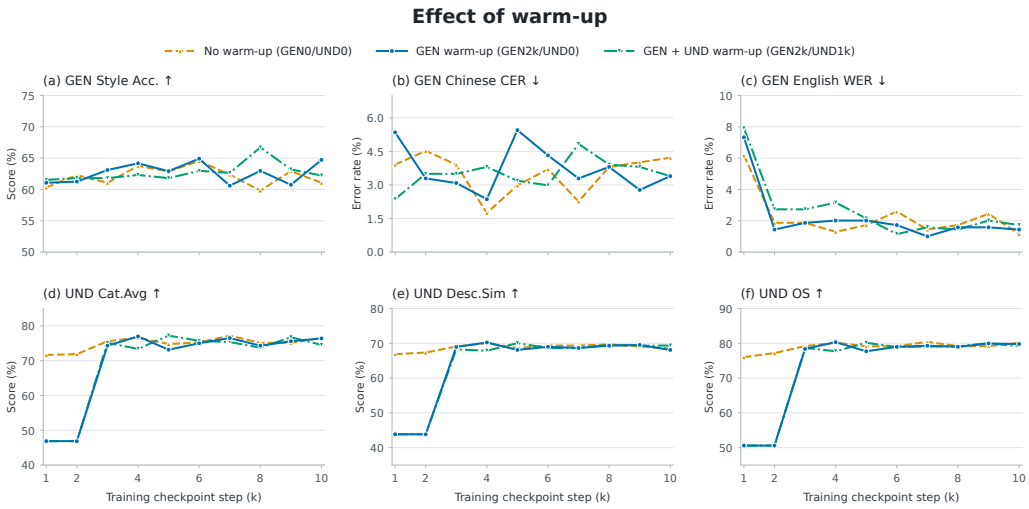

Figure S3: Joint-SFT trajectories through 10k training steps.

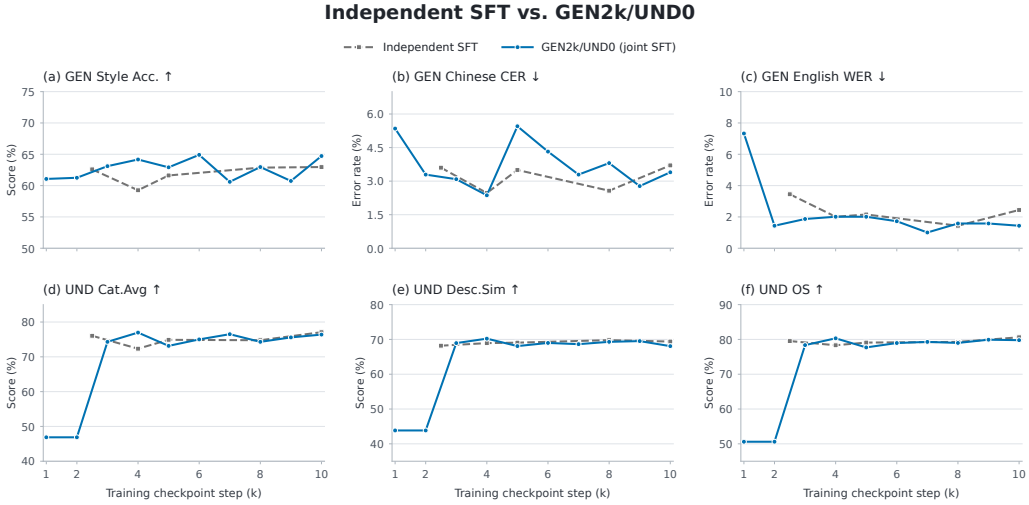

Figure S4: Comparison of Independent SFT (A1) and `GEN2k/UND0` through 10k training steps.

Results and GRPO initialization. As shown in Figures S3 and S4, `GEN2k/UND0` achieves the highest GEN Style Acc. among the four configurations at 10k steps (64.72%). Compared with Independent SFT, it also reduces Chinese CER from 3.70% to 3.40% and English WER from 2.44% to 1.44%, while its UND OS is slightly lower (79.78% versus 80.61%).

We therefore initialize CycleGRPO from `GEN2k/UND0` at 10k, prioritizing stronger generation control while retaining competitive profile-recovery performance.