ONE PROMPT, MANY VOICES: BENCHMARKING AND IMPROVING TIMBRE DIVERSITY IN INSTRUCTION-FOLLOWING VOICE DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Voice design requires generating natural speech that adheres to textual descriptions while providing diverse timbre candidates for applications like character casting. Since natural language descriptions constrain perceptual attributes without dictating a unique speaker identity, multiple distinct voices can satisfy a single instruction. However, current research primarily emphasizes instruction following and overlooks timbre diversity, leaving the generation of diverse yet compliant voices underexplored. To address this gap, we introduce VoxDICE, a benchmark combining hierarchical voice descriptions and a real speech reference set to jointly assess instruction following, conditional timbre variation, and global speaker distribution alignment. Evaluations of current models on VoxDICE reveal a prevalent empirical tension between instruction-following and timbre diversity. To overcome this limitation, we propose a reinforcement learning approach based on group relative policy optimization. This method integrates a distinct timbre reward within generation groups alongside an instruction gate that restricts diversity bonuses to compliant candidates. Experiments on the Qwen3 TTS model demonstrate significant improvements in both timbre diversity and instruction following, establishing that these capabilities can be jointly enhanced. Demos are available at https://voxdice744.github.io/voxdice/

1 INTRODUCTION

Advances in large language models (Brown et al., 2020; Guo et al., 2025; Du et al., 2022) have accelerated the development of speech synthesis by improving generation quality and enabling control through natural language instructions. Voice design has consequently emerged as an important research direction that allows users to create target timbres matching specific character profiles for diverse multimedia applications (Hu et al., 2026a; Zhou et al., 2026; Shen et al., 2026). These applications require models to generate natural speech following the given description while simultaneously providing sufficiently distinct candidate timbres. Practical tasks like character casting inherently involve selecting the most suitable voice from multiple diverse candidates that satisfy the same specification. If a model repeatedly generates similar timbres for the same description, the lack of meaningful alternatives for comparison substantially limits the practical value of voice design.

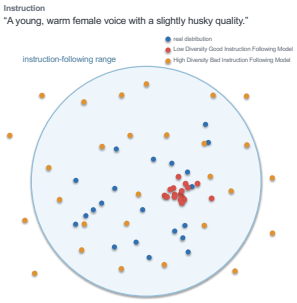

Figure 1: Schematic of weaker instruction following (orange) and low diversity (red). Blue points represent real voices; the shaded region denotes the range consistent with the same instruction.

Natural language voice descriptions typically specify only a subset of perceptual attributes and leave speaker identity underdetermined (Leng et al., 2024; Shimizu et al., 2024). For example, a description of a young and warm female voice can be satisfied by numerous perceptually distinct voices. Because the resulting mapping is inherently broad, an effective voice design model must be capable of generating a wide spectrum of distinct speaker identities that all faithfully comply with a single fixed instruction.

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.

Despite this inherent need for multiple valid realizations, existing voice design research has primarily emphasized instruction following, with comparatively limited attention to timbre diversity across repeated generations under the same description (Hu et al., 2026a; Shen et al., 2026). Consequently, progress in matching requested vocal attributes does not necessarily translate into a broader range of distinct candidate voices. Figure 1 illustrates this distinction through two failure modes. A model may follow the instruction closely while repeatedly producing similar timbres, or generate more varied timbres that deviate from the requested attributes. These failure modes highlight the need to assess and improve timbre diversity alongside instruction following, rather than treating either capability as sufficient on its own. However, systematic evaluation and explicit optimization of diverse yet compliant voice generation remain underexplored, leaving a gap between the capabilities emphasized in current research and the range of alternatives required in practical voice design (Chen et al., 2026; Huang et al., 2025).

To address this challenge, we introduce VoxDICE, a benchmark designed to evaluate timbre diversity and instruction consistency in voice design. We first construct a reference set comprising a maximally diverse collection of real human speakers paired with hierarchical voice descriptions. Utilizing this reference set, we formulate metrics to evaluate diversity across local condition scale and global distribution scale. Rigorous validation confirms that these proposed metrics effectively capture the nuanced interplay between timbre variation and semantic alignment, serving as a reliable proxy for human perceptual evaluation. Extensive evaluations on VoxDICE reveal a fundamental limitation in current voice design models. Existing approaches consistently exhibit a severe compromise, demonstrating either high timbre diversity with poor instruction adherence or strong instruction following with severely restricted diversity.

To enhance timbre diversity while maintaining instruction-following capabilities, we design a reward mechanism based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024). This mechanism integrates an inter-group diversity reward with an instruction-following gate. The diversity reward encourages the generation of distinct speaker embeddings within the same instruction group to promote acoustic variety. The instruction-following gate regulates this process by ensuring that the diversity bonus is exclusively awarded to candidates that strictly adhere to the given prompt. Empirical results demonstrate that our proposed reinforcement learning (RL) method effectively improves multi-scale diversity and simultaneously enhances the instruction-following capability of the model.

Our main contributions are as follows:

- We introduce VoxDICE, a comprehensive benchmark for voice design. Within this benchmark, we propose novel metrics at two distinct scales to evaluate timbre diversity, and empirically validate their effectiveness.
- Through extensive experiments, we reveal a prevalent empirical tension between instruction-following capability and timbre diversity among existing Voice Design systems.
- We propose a GRPO-based reward mechanism and demonstrate its effectiveness in addressing this limitation, successfully improving both timbre diversity and instruction-following performance.

2 Related Work

2.1 Natural-Language-Guided Voice Design

Natural-language-guided speech synthesis has progressed from controlling speaking style to designing speaker identities through textual descriptions. Early approaches such as PromptTTS (Guo et al., 2023) and InstructTTS (Yang et al., 2024) established text- based control over speech attributes, while VoxInstruct (Zhou et al., 2024) extended instruction-guided generation to multilingual and expressive speech. More recent systems, including VoiceSculptor (Hu et al., 2026b), Qwen3- TTS (Hu et al., 2026a), and MOSS-VoiceGenerator (Huang et al., 2026), support the creation of speaker timbres from natural- language descriptions. CapTalk (Su et al., 2026) further unifies caption-conditioned voice design for single-utterance and dialogue speech generation. These advances primarily emphasize speech quality and the correspondence between generated speech and

requested vocal characteristics. Systematic evaluation and explicit optimization of timbre diversity across repeated generations under a fixed description, however, remain largely underexplored.

2.2 Evaluating Diversity in Generative Models

Generative diversity evaluation in natural language processing and computer vision considers both variation under identical conditions (Yang et al., 2019; Mao et al., 2019) and global coverage of a reference distribution, since neither alone captures both conditional repetition and distributional mode dropping. To quantify local variation, similarity-based measures such as the Vendi Score (Friedman & Dieng, 2022) characterize redundancy among generated representations. Similarly, approaches like Palette (Saharia et al., 2022) compute the average embedding distance to evaluate diversity among generated samples. On the global scale, distributional approaches assess the coverage of real data using precision and recall, density and coverage, and retrieval-based measures (Kynkäänniemi et al., 2019; Naeem et al., 2020; Dombrowski et al., 2025). In speech synthesis, Yang et al. (2026) introduce ProsodyEval and a weighted edit-distance metric over discrete speech tokens to assess prosodic diversity and its agreement with human judgments. This work focuses on prosodic variation rather than variation in speaker identity.

The work most closely related to ours is MINT-Bench (Chen et al., 2026), which measures timbre diversity through the mean pairwise speaker distance among instruction-valid candidates weighted by the valid-candidate rate. However, this evaluation leaves two critical questions unresolved. First, it does not assess distributional diversity relative to a real-speaker reference set. Separation among candidates under individual conditions does not establish whether the generated voices collectively cover a broad range of speaker characteristics or repeatedly occupy a restricted region of the speaker space. Second, it lacks direct perceptual validation of its timbre-diversity score. The human-agreement analysis in MINT-Bench evaluates instruction-following judgments, leaving the correspondence between the diversity score and perceived timbre variation unestablished. VoxDICE addresses these gaps by combining conditional variation with global speaker-distribution alignment and coverage, validating the former against human timbre judgments and the latter through speaker-count experiments with a fixed sample count.

2.3 RL for Quality-Diversity Optimization

Reward-based post-training improves generation quality and controllability. However, pointwise quality rewards fail to distinguish novel valid candidates from repeated ones, often leaving mode collapse unaddressed. To mitigate this, recent studies in text and visual generation (Li et al., 2025; Lanchantin et al., 2025; Yuan et al., 2026; Liu et al., 2026; Li et al., 2026; Onoda et al., 2026) have introduced explicit diversity signals or group-level objectives, typically by coupling distribution-level diversity rewards with quality thresholds. However, these representations are tailored for text or visual modalities and do not address continuous timbre variation under natural-language vocal constraints. Inspired by these methods, we implement an algorithm combining a quality gate with an intra-group diversity reward to simultaneously enhance instruction-following capability and diversity in voice design.

3 VoxDICE: Benchmark Construction and Diversity Evaluation

3.1 Two Scales of Timbre Diversity

To systematically evaluate the timbre diversity of Voice Design models, we introduce metrics at two complementary scales. **Within-condition diversity** quantifies timbre variation among utterances generated from identical instructions and synthesis texts. **Global distribution diversity** assesses the alignment and coverage of generated timbres relative to an empirical real-speech reference distribution. Evaluating both dimensions prevents misleading conclusions because diverse outputs under individual instructions might collectively occupy an unrealistic region of the timbre space, whereas broad global coverage could conceal repetition within specific conditions.

3.2 Reference Corpus and Instruction Construction

3.2.1 Reference Corpus Construction

We use the Emilia dataset (He et al., 2024) as the data source as it contains speech from diverse recording scenarios and a wide range of emotional expressions, making it suitable for representing real-world speech variation. To guarantee high acoustic clarity and expressive richness within the reference distribution, we implement a rigorous filtering process. We retain only utterances that achieve arousal and dominance scores exceeding 0.8 estimated by dvatoolkit and yield a pinyin character error rate below 0.1 between the Qwen3-ASR (Shi et al., 2026) outputs and the ground truth transcripts. After filtering, we promote broad speaker coverage by sampling at most one segment from each original audio file. An internal classifier assigns each utterance to a male, female, or child category. We sample equal numbers of male and female utterances and retain a subset of child speech to preserve age diversity. The resulting 10,000 utterances form the empirical reference distribution.

3.2.2 Instruction Annotation

Following InstructTTSEval (Huang et al., 2025), we categorize voice design instructions into three scales comprising Acoustic-Parameter Specification (APS), Descriptive-Style Directive (DSD), and Role-Play (RP). We first employ Gemini-3.1-Pro-Preview to extract APS instructions from the audio and subsequently utilize GPT-4o to derive the corresponding DSD and RP instructions. We adopt the speech-captioning and instruction-generation prompt templates from InstructTTSEval; the templates are reproduced in Appendix A.6.

3.3 Diversity Metrics

3.3.1 Within-Condition Diversity

We employ speaker embeddings extracted via CAM++ (Wang et al., 2023) as the intermediate representation for diversity evaluation. Given an instruction c and a synthesis text t, we generate K utterances and denote the embedding of the k-th utterance as $z_{c,t,k} = f_{\mathrm{CAM++}}(x_{c,t,k})$. We quantify the intra-condition variation using two complementary metrics.

Mean Pairwise Cosine Distance (MPCD) This metric computes the average spatial distance among speaker embeddings within a single generation set. Let $\hat{z}_{c,t,i} = z_{c,t,i}/\|z_{c,t,i}\|_2$ denote the normalized speaker embedding. We formulate it as

$$\mathrm{MPCD}(c,t) = \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} \left(1 - \hat{z}_{c,t,i}^{\top} \hat{z}_{c,t,j}\right). \tag{1}$$

A higher value reflects a more pronounced average timbre separation under the given condition.

Vendi Score (VS) We additionally compute Vendi Scores of the generated embeddings (Friedman & Dieng, 2022). This metric formally evaluates the effective number of distinct speaker representations by measuring the von Neumann entropy of the similarity matrix. We construct the pairwise cosine similarity matrix $S^{(c,t)}$ with elements

$$S_{ij}^{(c,t)} = \hat{z}_{c,t,i}^{\top} \hat{z}_{c,t,j}. \tag{2}$$

Let $\lambda_1, \ldots, \lambda_K$ denote the eigenvalues of $S^{(c,t)}/K$, which sum to one. The score is given by

$$\mathrm{VS}(c,t) = \exp\left(-\sum_{k=1}^{K} \lambda_k \log \lambda_k\right). \tag{3}$$

Unlike MPCD, which measures average pairwise separation and may overestimate diversity under mode collapse, VS captures the effective number of distinct speaker representations through the embedding set's intrinsic dimensionality.

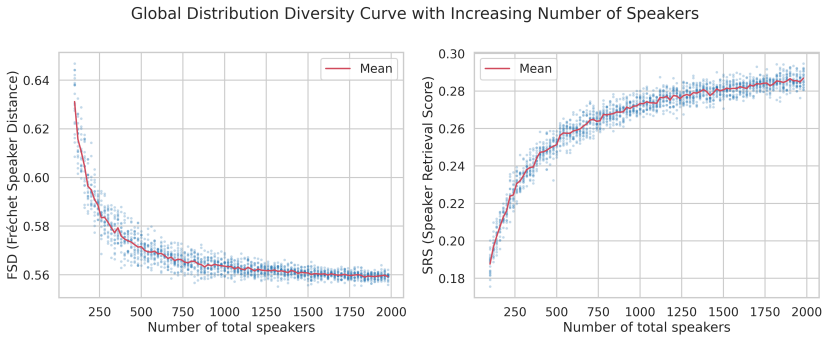

Figure 2: Global distribution diversity metrics as the number of speakers increases while the total number of utterances remains fixed. Each blue point represents one run, and the red curve denotes the mean over repetitions. FSD decreases as speaker diversity increases, whereas SRS increases.

3.3.2 Global Distribution Diversity

Let $\mathcal{R} = \{z_i^r\}_{i=1}^{N_r}$ and $\mathcal{G} = \{z_j^g\}_{j=1}^{n_g}$ denote the reference and generated embedding sets, respectively. In our settings, $n_g = N_r = 10{,}000$. Global diversity evaluates the distribution-level difference between these sets using two complementary metrics.

Fréchet Speaker Distance (FSD). We compute the Fréchet distance between the two speaker embedding distributions. The FSD metric is conceptually analogous to the widely used Fréchet Inception Distance (Heusel et al., 2017), with the key difference being the replacement of image-based Inception features with our speaker embeddings. Let (μ_r, Σ_r) and (μ_g, Σ_g) be their empirical means and covariance matrices. Then

$$\mathrm{FSD}(\mathcal{R}, \mathcal{G}) = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\right). \tag{4}$$

Lower FSD indicates closer global first- and second-order statistics between the generated and real speaker distributions.

Speaker Retrieval Score (SRS). Inspired by BeyondFID (Dombrowski et al., 2025), we employ the SRS to measure the coverage of generated speaker characteristics. We replace the image representations used by the original Image Retrieval Score with speaker embeddings and retain the same nearest-neighbor retrieval rule. Let $s(\cdot, \cdot)$ denote cosine similarity and define

$$\pi(j) = \arg \max_{i \in \{1, \ldots, N_r\}} s(z_j^g, z_i^r). \tag{5}$$

The intuitive finite-sample coverage is

$$\mathrm{SRS}_\alpha = \frac{1}{N_r} \left| \{\pi(j) : j = 1, \ldots, n_g\} \right|. \tag{6}$$

Unlike FSD, which measures global statistical alignment, SRS measures the fraction of reference support retrieved by the generated set, so higher values indicate broader speaker coverage. Throughout the paper, SRS without a subscript denotes the adjusted asymptotic score $\mathrm{SRS}_\infty^{\mathrm{adjusted}}$. The infinite-sample extrapolation and measurement-gap correction used to compute this reported score are described in Appendix A.1.

3.3.3 Global Metric Validation through Speaker Insertion

To demonstrate the effectiveness of the proposed global distribution diversity metrics, we artificially construct data with varying speaker diversity to compare against a reference set. Specifically, we sample varying numbers of speakers ranging from 20 to 1991 from the AISHELL-2 dataset (Du et al., 2018) while fixing the total number of utterances at 10000. The utterance allocation is strictly balanced, ensuring the number of utterances contributed by each speaker differs by at most one. For

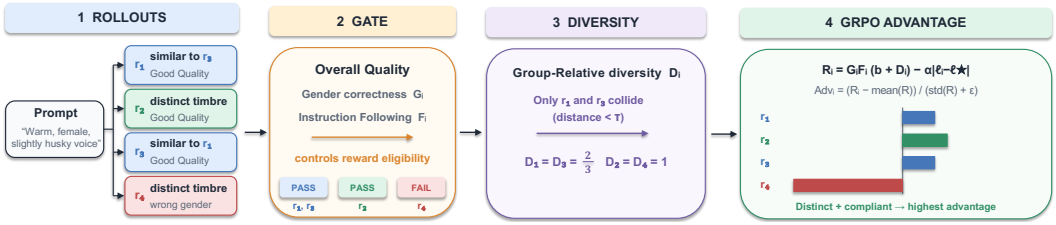

Figure 3: Overview of the proposed GRPO procedure. For each instruction-text pair, the policy samples a rollout group, applies gender and instruction-following gates, and computes group-relative timbre diversity from speaker-embedding collisions. The resulting reward is normalized within the group to form the GRPO advantage. Here, r_1 and r_3 collide, r_2 is compliant and distinct, and r_4 is distinct but fails the gender gate. Consequently, r_2 receives the highest advantage, whereas r_4 receives no gated quality or diversity reward.

each speaker count, we perform 20 repeated sampling iterations and calculate the average FSD and SRS values.

As illustrated in Figure 2, an increase in the number of speakers leads to a clear rise in SRS and a decline in FSD. Notably, this trend is nonlinear since the rate of change gradually slows down as the speaker count increases. This deceleration reflects a natural saturation effect in the acoustic embedding space as the global timbre diversity approaches the inherent upper limit of the AISHELL-2 dataset. To quantify this relationship, we compute the Spearman correlations of ρ, which are well suited for evaluating nonlinear monotonic trends. The correlation coefficient is -0.88 for FSD and 0.95 for SRS (both $p < 10^{-50}$). This confirms that both metrics effectively capture the progression of global timbre diversity and validates that speaker embeddings serve as a reliable intermediate representation for such assessment.

4 Instruction-Gated RL for Timbre Diversity

To enhance timbre diversity while maintaining instruction adherence, we design a tailored reward for GRPO. As summarized in Figure 3, each optimization step consists of four stages. First, the current policy generates a group of candidate utterances from the same voice instruction and synthesis text. Second, each candidate is checked for gender correctness and instruction following. Third, the candidates are compared in speaker-embedding space to identify near-duplicate timbres within the group. Finally, the gated quality-diversity reward is standardized within the rollout group to obtain the GRPO advantage.

4.1 Optimization Setup and Reward Formulation

For an instruction-text pair x, the policy samples a group $\{r_i\}_{i=1}^{K}$. We evaluate each candidate using two binary eligibility indicators and a group-relative diversity score. Let $G_i \in \{0, 1\}$ denote gender correctness, $F_i \in \{0, 1\}$ denote instruction-following acceptance, and $D_i \in [0, 1]$ denote timbre distinctiveness within the group. The reward is

$$R_i = G_i F_i \big(b + D_i\big) - \alpha |\ell_i - \ell^\star|, \tag{7}$$

where b is the base reward for an eligible candidate and the final term penalizes deviations of the generated token length ℓ_i from the reference length $\ell^\star$. Thus, the product $G_i F_i$ gates both the base reward and the diversity bonus. A distinct candidate that violates either constraint cannot obtain a positive reward merely by moving away from the other speakers. The length term remains outside the gate so that malformed sequence lengths are penalized regardless of candidate eligibility.

GRPO converts these rewards into group-relative advantages without a learned value model:

$$\widehat{A}_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon}, \qquad \mu_R = \frac{1}{K} \sum_{j=1}^{K} R_j, \tag{8}$$

where σ_R is the within-group reward standard deviation and ϵ is a numerical stabilizer. The policy is then updated using the GRPO surrogate objective together with a KL penalty to the reference policy.

4.2 Instruction-Following Gating

We employ Qwen3-Omni (Xu et al., 2025) to evaluate the instruction following capability of the generated audio. To mitigate the hallucinations frequently observed in large audio-language models, we additionally introduce an internal gender discriminator for gender gating. Specifically, we define the gender label space as $\mathcal{G} = \{\text{male}, \text{female}, \text{uncertain}\}$. The instruction text in the training data is first annotated by Gemini-3.1-Pro-Preview to obtain a target label $g \in \mathcal{G}$, and the internal gender discriminator subsequently verifies whether the generated audio aligns with g. Furthermore, to maximize the discriminative capability of Qwen3-Omni, we proactively reformulate the descriptions of DSD and RP into concrete APS-style evaluation instructions. This reformulation is exclusively applied during the scoring phase and preserves the original policy input. The APS-reformulation prompt is provided in Appendix A.6.4.

4.3 Collision-Based Timbre Diversity Reward

Let r_i denote the speaker embedding of the i-th generated candidate, and let d_{ij} be the cosine distance between r_i and r_j. A pair is considered a timbre collision when $d_{ij} < \tau$. We define

$$D_i = \frac{1}{K-1} \sum_{j \neq i} \mathbb{I}[d_{ij} \geq \tau] \,, \tag{9}$$

namely, the proportion of the other candidates that do not collide with r_i. The comparison is local to candidates generated from the same condition, so the signal specifically targets within-condition repetition. Diversity is computed for the complete rollout group, but Equation 7 makes it reward-eligible only when both quality gates pass.

4.4 Implementation Details

We utilize a private multi-speaker dataset as the training data source. The annotation procedure strictly follows InstructTTSEval. This dataset ensures no overlap with the Emilia evaluation set. During training, we unfreeze all parameters of Qwen3-TTS-12Hz-1.7B-VoiceDesign except the tokenizer. For the GRPO configuration, we employ a batch size of 8 with 8 rollouts. The hyperparameters include a learning rate of 10^{-6}, a base reward $b = 3$, a length penalty coefficient $\alpha = 0.01$, a KL penalty coefficient $\beta_{\text{KL}} = 0.01$, a sampling temperature of 1.0, and a collision threshold of 0.7.

5 Experiments

5.1 Experimental Setup

We evaluate seven voice design models, including VoiceSculptor (Hu et al., 2026b), Qwen3-TTS-12Hz-1.7B-VoiceDesign (Hu et al., 2026a), VoxCPM2 (Zhou et al., 2026), Ming-omni-tts-0.5B[1], MiMo-Audio-7B-Instruct (Xiaomi, 2025), MOSS-VoiceGenerator (Huang et al., 2026), and FireRedTTS3-Instruct (Shen et al., 2026). Additionally, we evaluate the Qwen3-TTS model trained with our proposed RL algorithm, hereafter referred to as Qwen3-TTS-GRPO. We do not evaluate closed-source models such as Gemini-2.5-Pro-TTS and GPT-4o-mini-TTS, as their voice design workflows require specifying a reference timbre as a starting point, which is inconsistent with our goal of assessing timbre diversity without a predefined timbre reference.

For each model, we use its official inference settings to generate $K = 10$ utterances via repeated sampling for every benchmark case across different instruction scales, with the instruction and synthesis text held fixed. To ensure randomness, we modify the random seed before each generation.

Utilizing the generated data, we compute the MPCD and VS at the case level, and subsequently calculate the mean and standard deviation across different instruction scales for each model. To assess global distribution diversity, we sample one generated utterance per case to construct a generated set that matches the size of the reference set, which is then used to compute the FSD and SRS. To evaluate the instruction following capabilities of the models, we randomly sample 1000 cases

[1] https://github.com/inclusionAI/Ming-omni-tts

Table 1: Within-condition timbre diversity across three prompt scales. All entries report mean ± standard deviation across cases. Average reports the unweighted mean of three scale-specific means, not a pooled estimate. Higher is better for both metrics; the best mean in each column is bolded.

Prompt scale	APS		DSD		RP		Average	
Models	MPCD ↑	VS ↑	MPCD ↑	VS ↑	MPCD ↑	VS ↑	MPCD ↑	VS ↑
VoiceSculptor	0.284 ± 0.100	2.806 ± 0.734	0.327 ± 0.085	3.273 ± 0.669	0.340 ± 0.085	3.373 ± 0.674	0.317	3.151
Qwen3-TTS-12Hz-1.7B-VoiceDesign	0.282 ± 0.068	2.913 ± 0.557	0.359 ± 0.106	3.524 ± 0.898	0.359 ± 0.117	3.520 ± 0.994	0.333	3.319
VoxCPM2	0.440 ± 0.105	4.390 ± 1.031	0.554 ± 0.153	5.542 ± 1.623	0.555 ± 0.127	5.517 ± 1.351	0.516	5.150
Ming-omni-tts-0.5B	0.384 ± 0.104	3.784 ± 0.934	0.416 ± 0.112	4.094 ± 1.034	0.406 ± 0.103	3.986 ± 0.932	0.402	3.955
MiMo-Audio-7B-Instruct	$\mathbf{0.683} \pm 0.085$	$\mathbf{6.953} \pm 0.940$	$\mathbf{0.689} \pm 0.085$	$\mathbf{7.028} \pm 0.933$	$\mathbf{0.691} \pm 0.084$	$\mathbf{7.046} \pm 0.929$	**0.688**	**7.009**
MOSS-VoiceGenerator	0.639 ± 0.116	6.382 ± 1.219	0.628 ± 0.116	6.273 ± 1.220	0.595 ± 0.103	5.921 ± 1.080	0.621	6.192
FireRedTTS3-Instruct	0.223 ± 0.080	2.434 ± 0.561	0.249 ± 0.093	2.584 ± 0.622	0.256 ± 0.099	2.618 ± 0.645	0.243	2.545
Qwen3-TTS-GRPO	0.345 ± 0.075	3.456 ± 0.654	0.420 ± 0.094	4.085 ± 0.846	0.416 ± 0.105	4.040 ± 0.928	0.394	3.860

Table 2: Global distribution diversity and instruction following across three prompt scales. Average reports the unweighted arithmetic mean of each metric across APS, DSD, and RP. The best value in each column is bolded.

Prompt scale	APS			DSD			RP			Average		
Models	FSD ↓	SRS ↑	IFR (%) ↑	FSD ↓	SRS ↑	IFR (%) ↑	FSD ↓	SRS ↑	IFR (%) ↑	FSD ↓	SRS ↑	IFR (%) ↑
VoiceSculptor	0.734	0.084	43.6	0.741	0.105	64.3	0.736	0.107	40.3	0.737	0.099	49.4
Qwen3-TTS-12Hz-1.7B-VoiceDesign	0.748	0.062	77.9	0.674	0.094	81.1	0.654	0.094	50.0	0.692	0.083	69.7
VoxCPM2	0.469	0.151	61.5	0.307	0.313	76.6	0.343	0.260	41.1	0.373	0.241	59.7
Ming-omni-tts-0.5B	0.423	0.221	76.8	0.423	0.239	81.0	0.445	0.222	44.9	0.430	0.227	67.6
MiMo-Audio-7B-Instruct	**0.245**	**0.393**	46.4	**0.219**	**0.424**	63.9	**0.222**	**0.415**	42.1	**0.229**	**0.411**	50.8
MOSS-VoiceGenerator	0.424	0.332	52.7	0.334	0.384	73.7	0.359	0.307	46.6	0.372	0.341	57.7
FireRedTTS3-Instruct	0.651	0.094	76.8	0.649	0.097	82.2	0.631	0.101	**50.6**	0.644	0.097	69.9
Qwen3-TTS-GRPO	0.679	0.088	**83.3**	0.641	0.111	**86.6**	0.618	0.112	48.4	0.646	0.104	**72.8**

and design prompts across three scales for each case, employing Gemini-3.1-Pro-Preview[2] as the evaluator. Ultimately, we adopt the instruction following rate (IFR) as the evaluation metric.

The LLM-judge prompt and dimension-specific criteria are in AppendixA.6.5, with human–LLM agreement results in AppendixA.2. We use different models for training-time reward estimation and final evaluation, as open-source large audio-language models remain prone to hallucinations and are insufficiently reliable as evaluation judges.

5.2 Benchmark Findings

Tables 1 and 2 show that strong instruction following does not necessarily imply diverse voice generation. Among the unmodified baselines, Qwen3-TTS and FireRedTTS3-Instruct lead in instruction following but exhibit limited diversity at both scales. Conversely, MiMo-Audio, MOSS-VoiceGenerator, and VoxCPM2 produce more diverse timbres but follow instructions less reliably. This contrast highlights why neither capability alone is sufficient for evaluating voice design. Ming-omni-TTS offers a more balanced profile, combining near-leading instruction following with middle-ranking diversity, whereas VoiceSculptor performs poorly on both capabilities.

Within-condition variation and global coverage also capture distinct behaviors. Under APS, VoxCPM2 produces more varied timbres for the same instruction and synthesis text than Ming-omni-TTS, yet Ming-omni-TTS better covers and aligns with the reference speaker distribution. This ranking reversal shows that locally varied outputs need not collectively span a broader timbre space, supporting the need to evaluate diversity at both scales.

5.3 Effectiveness of GRPO

Instruction-gated GRPO improves all diversity metrics over the original Qwen3-TTS checkpoint across all scales, indicating gains in both within-condition variation and global distribution diversity. Instruction following improves on APS and DSD, with a slight decline on RP. Overall, average IFR

[2]Due to budget constraints, we do not evaluate all cases. Furthermore, as Gemini-2.5-Pro is scheduled for deprecation, we utilize Gemini-3.1-Pro-Preview as the LLM judge for instruction following evaluation to maintain experimental reproducibility, which differs from InstructTTSEval.

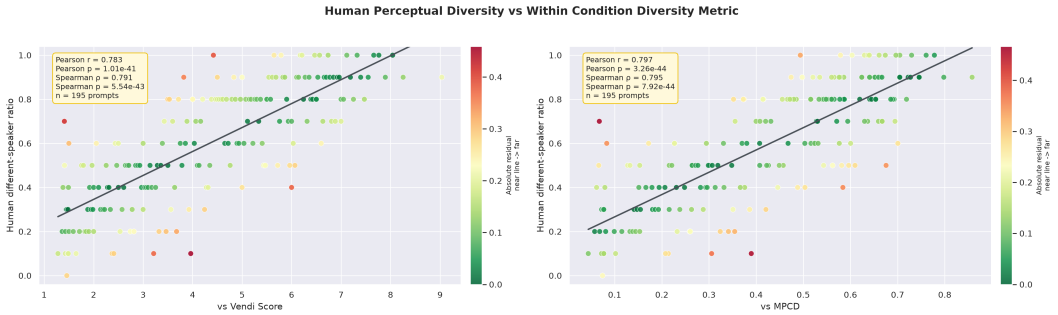

Figure 4: Alignment between human judgments of within-condition timbre diversity and automatic metrics across audio groups. Both VS and MPCD exhibit strong positive correlations with human scores. The color bars to the right of each panel indicate absolute residuals from the fitted regression line, with cooler (green) colors denoting smaller and warmer (orange/red) colors denoting larger.

rises from 69.7% to 72.8%, a gain of 3.1 percentage points and the highest score among the evaluated models. These results demonstrate that timbre diversity and instruction following are not mutually exclusive: our RL approach improves diversity and aggregate instruction following simultaneously.

The evaluation in AppendixA.3 finds no accompanying degradation in speech quality, naturalness, or intelligibility, further supporting our approach. Ablation experiments on the gender gate, instruction-following gate, diversity bonus, and scoring-time APS reformulation are in AppendixA.4.

5.4 Perceptual Validation of Within-Condition Diversity Metrics

We assess whether the within-condition diversity metrics, VS and MPCD, align with human judgments. Rather than scoring an entire audio group, we quantify perceived timbre diversity using pairwise judgments. We define the human diversity score as $H = 1 - \widehat{p}_{\text{same}}$, where $\widehat{p}_{\text{same}}$ is the estimated probability that two sampled utterances from the same group share the same timbre.

Human evaluation protocol. To ensure broad coverage, we stratify audio groups by VS into 13 bins, sampling 15 groups per bin (195 total). For each group, three participants judge 10 random audio pairs for timbre consistency (annotation interface in Appendix A.5). We recruited 30 participants,[3] yielding 5,850 judgments across 1,950 pairs. Aggregating these by majority vote, H is calculated as the proportion of pairs labeled as having different timbres.

Results. As shown in Figure 4, VS achieves Pearson and Spearman correlations of $r = 0.783$ and $\rho = 0.791$ with human scores; MPCD achieves $r = 0.797$ and $\rho = 0.795$. Both show strong, statistically significant positive associations ($p < 10^{-40}$), validating their perceptual reliability for within-condition timbre diversity assessment. Together with the speaker insertion experiment, these results demonstrate that CAM++ speaker embeddings faithfully capture speaker identity distinctions, supporting their use as a reliable proxy for speaker diversity measurement.

6 Conclusion

Voice design requires diverse speaker identities that satisfy the same description. We introduced VoxDICE to evaluate this capability through within-condition timbre variation and global speaker-distribution alignment and coverage, alongside instruction following. Evaluations of seven models show that strong instruction adherence often coexists with limited timbre diversity, while greater diversity can accompany weaker adherence. To address this tension, we proposed instruction-gated GRPO, which rewards distinct timbres only when candidates meet the instruction constraints. On Qwen3-TTS, this approach improves diversity at both scales and raises average instruction following, demonstrating that the two capabilities can improve jointly. These findings support evaluating and optimizing timbre diversity and instruction adherence together in voice design.

[3]Each participant spent approximately 40 minutes on the study and received RMB 30 in compensation.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for developing theoretical models or conceptual frameworks, proposing or refining hypotheses, or designing or providing feedback on research methodology or experiments. Formulating mathematical claims, providing critical ingredients for mathematical proofs, and assisting with proof writing are not applicable to this work. Additionally, we used generative AI tools for improving the language and readability of the manuscript, drafting parts of the manuscript, searching for, summarizing, or analyzing existing literature, creating or modifying scientific figures, and suggesting manuscript structure, titles, or keywords. We have reviewed all AI-assisted work. Specifically, we checked the correctness of LLM-generated instruction annotations, reviewed and tested AI-assisted code, and cross-checked AI-assisted analyses and interpretations of results. AI-assisted translations, figures, and manuscript text were also reviewed by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the benchmark construction and evaluation metrics in Section 3, the training implementation details in Section 4.4, and the evaluation setup in Section 5.1. The prompt templates used for instruction annotation and evaluation are provided in Appendix A.6. We will publicly release the benchmark, evaluation code, and trained model checkpoints immediately upon acceptance of the paper to facilitate independent evaluation and further research.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Huakang Chen, Jingbin Hu, Liumeng Xue, Qirui Zhan, Wenhao Li, Guobin Ma, Hanke Xie, Dake Guo, Linhan Ma, Yuepeng Jiang, et al. Mint-bench: A comprehensive multilingual benchmark for instruction-following text-to-speech. *arXiv preprint arXiv:2604.17958*, 2026.

Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3029–3039. IEEE, 2025.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890. IEEE, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026a.

Jingbin Hu, Huakang Chen, Linhan Ma, Dake Guo, Qirui Zhan, Wenhao Li, Haoyu Zhang, Kangxiang Xia, Ziyu Zhang, Wenjie Tian, et al. Voicesculptor: Your voice, designed by you. *arXiv preprint arXiv:2601.10629*, 2026b.

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.

Kexin Huang, Liwei Fan, Botian Jiang, Yaozhou Jiang, Qian Tu, Jie Zhu, Yuqian Zhang, Yiwei Zhao, Chenchen Yang, Zhaoye Fei, et al. Moss-voicegenerator: Create realistic voices with natural language descriptions. *arXiv preprint arXiv:2603.28086*, 2026.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

Yichong Leng, Zhifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, Kaitao Song, Lei He, et al. Prompttts 2: Describing and generating voices with text prompt. In *International Conference on Learning Representations*, volume 2024, pp. 57672–57688, 2024.

Jiaming Li, Chenyu Zhu, Nanxi Yi, Youjun Bao, Li Sun, Quanying Lv, Xiang Fang, Daizong Liu, Jianjun Li, Kun He, et al. Tmpo: Trajectory matching policy optimization for diverse and efficient diffusion alignment. *arXiv preprint arXiv:2605.10983*, 2026.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025.

Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. Diversegrpo: Mitigating mode collapse in image generation via diversity-aware grpo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1864–1873, 2026.

Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1429–1437. IEEE, 2019.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020.

Ku Onoda, Paavo Parmas, Hiroki Furuta, Soichiro Nishimori, Yuta Oshima, Shohei Taniguchi, and Yutaka Matsuo. Multi-axis max@ k reinforcement learning for representative diversity in text-to-image generation. *arXiv preprint arXiv:2607.14962*, 2026.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Feiyu Shen, Kun Xie, Yichen Wu, Ziqi Dai, Yichen Han, Junjie Li, Xuelong Geng, Fenglong Xie, Lei Xie, Xu Tang, et al. Fireredtts3: Unified speech generation and editing with semantically enriched speech representations. *arXiv preprint arXiv:2608.17492*, 2026.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12672–12676. IEEE, 2024.

Xiaosu Su, Zihan Sun, Peilei Jia, and Jun Gao. Captalk: Unified voice design for single-utterance and dialogue speech generation. *arXiv preprint arXiv:2604.08363*, 2026.

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*, 2023.

LCT Xiaomi. Mimo-audio: Audio language models are few-shot learners. *CoRR, abs/2512.23808*, 2025.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*, 2019.

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.

Yifan Yang, Bing Han, Hui Wang, Long Zhou, Wei Wang, Mingyu Cui, Xu Tan, and Xie Chen. Measuring prosody diversity in zero-shot tts: A new metric, benchmark, and exploration. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17252–17256. IEEE, 2026.

Jiayi Yuan, Hangoo Kang, James Jihao Liu, Yejin Choi, Vikram Iyer, Liwei Jiang, and Natasha Jaques. Forty shades of blue: Quality-diversity alignment via mode-conditioned reinforcement learning. *arXiv preprint arXiv:2609.14896*, 2026.

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.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, Jiaheng Wu, Ziyang Wang, Xudong Shen, Runchuan Ye, et al. Voxcpm2 technical report. *arXiv preprint arXiv:2606.06928*, 2026.

A Appendix

A.1 Detailed Computation of the Speaker Retrieval Score

This appendix gives the complete computation of the Speaker Retrieval Score used in the experiments. Throughout the paper, the unqualified notation SRS refers to the adjusted asymptotic score $\mathrm{SRS}_{\infty}^{\mathrm{adjusted}}$ defined below. We first describe the embedding and retrieval procedure, then derive the finite-sample coverage, occupancy-based extrapolation, confidence interval, and measurement-gap correction.

We compute the Speaker Retrieval Score using CAM++ speaker embeddings. Let

$$\mathcal{G} = \{z_j^g\}_{j=1}^{n_g} \quad \text{and} \quad \mathcal{R} = \{z_i^r\}_{i=1}^{N_r}$$

denote the generated and reference embedding sets, respectively. In our experiments, both sets contain 10,000 embeddings and each CAM++ embedding has dimension 192.

Embedding normalization and retrieval. Each embedding is independently ℓ_2-normalized:

$$\tilde{z} = \frac{z}{\max(\|z\|_2, \epsilon)}, \qquad \epsilon = 10^{-12}. \tag{10}$$

The similarity between two embeddings is cosine similarity:

$$s(z_a, z_b) = \tilde{z}_a^\top \tilde{z}_b. \tag{11}$$

For every generated embedding z_j^g, we retrieve the most similar reference embedding:

$$\pi(j) = \arg\max_{i \in \{1,\dots,N_r\}} s(z_j^g, z_i^r). \tag{12}$$

The retrieval operation is performed in chunks to avoid materializing the full $n_g \times N_r$ similarity matrix at once. The chunking does not change the resulting nearest-neighbor assignments.

Finite-sample SRS. Let

$$k = |\{\pi(j) : j = 1, \dots, n_g\}| \tag{13}$$

be the number of distinct reference embeddings retrieved by the generated set. The finite-sample retrieval coverage is

$$\mathrm{SRS}_\alpha = \frac{k}{N_r}. \tag{14}$$

This score measures the fraction of reference support covered by the observed generated samples. Multiple generated samples retrieving the same reference embedding count only once.

Occupancy model for infinite-sample extrapolation. The observed coverage is affected by the finite number of generated samples. To estimate the coverage attainable with unlimited generation, we use an occupancy model. We assume that the generator effectively samples uniformly from an unknown set of D reachable reference identities, where

$$k \le D \le N_r.$$

Given n_g generated samples, the probability of observing exactly k distinct retrieved identities is

$$P(k \mid D, n_g) = \left\{ {n_g \atop k} \right\} \frac{D!}{(D-k)! D^{n_g}}, \tag{15}$$

where $\left\{ {n_g \atop k} \right\}$ is a Stirling number of the second kind. It counts the number of ways to partition the n_g generated samples into k nonempty groups.

For numerical stability, the likelihood is evaluated in log space:

$$\log P(k \mid D, n_g) = \log \left\{ {n_g \atop k} \right\} + \log \Gamma(D+1) - \log \Gamma(D-k+1) - n_g \log D. \tag{16}$$

The Stirling numbers are computed using the recurrence

$$S(n,k) = kS(n-1,k) + S(n-1,k-1), \tag{17}$$

with $S(0,0) = 1$ and $S(n,0) = 0$ for $n > 0$.

The effective support size is estimated by maximum likelihood:

$$\widehat{D} = \arg\max_{D \in \{k, k+1, \dots, N_r\}} \log P(k \mid D, n_g). \tag{18}$$

The resulting infinite-sample retrieval score is

$$\mathrm{SRS}_\infty = \frac{\widehat{D}}{N_r}. \tag{19}$$

Thus, SRS_∞ estimates the fraction of the reference support that would eventually be covered if the generator were sampled indefinitely under the occupancy model.

Confidence interval. We obtain a likelihood-ratio confidence interval for $\widehat{D}$. A candidate value D is retained when

$$2\left[\log P(k \mid \widehat{D}, n_g) - \log P(k \mid D, n_g)\right] \le \chi^2_{1,0.95}, \tag{20}$$

where $\chi^2_{1,0.95} = 3.8414588$. If the resulting integer interval is $[D_{\mathrm{lo}}, D_{\mathrm{hi}}]$, the corresponding confidence interval for the unadjusted score is

$$\left[\frac{D_{\mathrm{lo}}}{N_r}, \frac{D_{\mathrm{hi}}}{N_r}\right]. \tag{21}$$

Measurement-gap correction. The CAM++ embedding space has finite resolution: even distinct reference speakers may retrieve the same nearest neighbor. This creates a measurement gap between the diversity represented in the data and the diversity measurable by the embedding extractor. We estimate this gap by leave-one-out self-retrieval on the reference set.

For each reference embedding, self-retrieval excludes the embedding itself:

$$\pi_{\mathrm{real}}(i) = \arg\max_{\ell \in \{1,\ldots,N_r\},\, \ell \neq i} s(z^r_i, z^r_\ell). \tag{22}$$

The number of distinct self-retrieved reference entries is

$$k_{\mathrm{real}} = \left|\{\pi_{\mathrm{real}}(i) : i = 1, \ldots, N_r\}\right|. \tag{23}$$

We then apply exactly the same occupancy MLE procedure using $n = N_r$ and $k = k_{\mathrm{real}}$:

$$\widehat{D}_{\mathrm{real}} = \arg\max_{D \in \{k_{\mathrm{real}},\ldots,N_r\}} \log P(k_{\mathrm{real}} \mid D, N_r). \tag{24}$$

The measurable coverage ceiling of the CAM++ representation is

$$\mathrm{SRS}^{\mathrm{real}}_{\infty} = \frac{\widehat{D}_{\mathrm{real}}}{N_r}. \tag{25}$$

Following the measurement-gap correction used in the original IRS formulation, the final adjusted score is

$$\mathrm{SRS}^{\mathrm{adjusted}}_{\infty} = \frac{\mathrm{SRS}_{\infty}}{\mathrm{SRS}^{\mathrm{real}}_{\infty}}. \tag{26}$$

The same reference set and the same CAM++ embedding space are used for both the generated-set retrieval and the self-retrieval baseline. The finite-sample score SRS_{α} is retained as an uncorrected diagnostic, while the measurement-gap correction is applied to the infinite-sample extrapolation used for the reported SRS.

The reported metric denoted by SRS throughout the paper is $\mathrm{SRS}^{\mathrm{adjusted}}_{\infty}$. The finite-sample quantity SRS_{α} and the unadjusted extrapolation SRS_{∞} are intermediate quantities used in its computation. In the implementation, the names IRS_{α}, IRS_{∞}, and $\mathrm{IRS}^{\mathrm{adjusted}}_{\infty}$ are retained as compatibility aliases for the corresponding SRS quantities.

A.2 LLM Judge Consistency

To assess the reliability of our automated instruction-following evaluation, we compare the LLM judge with human annotations across the three instruction scales: APS, DSD, and RP. We sample 300 instruction–audio pairs from each scale, yielding 900 cases in total. We recruit 10 annotators, each labeling 270 cases,[4] for a total of 2,700 human judgments. Each case receives three human annotations indicating whether the audio follows the instruction. We obtain the final human label by majority vote over these three annotations and compare it with the label produced by Gemini-3.1-Pro-Preview, the LLM judge used in our benchmark evaluation. Human–LLM agreement is measured as the percentage of cases for which the LLM label matches the majority-vote human label. The instruction-following annotation interface is shown in Figure 6 in Appendix A.5.

[4]Each annotator spent approximately one hour completing the task and received RMB 50 in compensation.

Table 3: Agreement between the LLM judge and majority-vote human annotations for instruction following. Each instruction scale contains 300 sampled instruction–audio pairs, with three human annotations per pair.

Instruction scale	APS	DSD	RP
Human–LLM agreement (%)	80.7	83.3	72.3

Table 3 shows relatively high agreement between human annotations and the LLM judge across all three instruction scales, with higher agreement on APS and DSD than on RP. These results support using the LLM judge as a practical proxy for human annotation in large-scale instruction-following evaluation, reducing the need for manual labeling. Nevertheless, the remaining disagreements, particularly on RP, indicate that automated judgments are not fully interchangeable with human labels and motivate retaining human checks for ambiguous cases.

A.3 Additional Evaluations after RL

To assess whether improvements in timbre diversity and instruction following come at the expense of speech naturalness, intelligibility, or audio quality, we additionally evaluate Qwen3-TTS-12Hz-1.7B-VoiceDesign checkpoint and Qwen3-TTS-GRPO, using DNSMOS, UTMOSv2, and pinyin CER. DNSMOS and UTMOSv2 serve as automatic indicators of speech quality and naturalness, while pinyin CER assesses intelligibility through transcription accuracy.

Evaluation procedure. For pinyin CER, we first transcribe the generated audio using Qwen3-ASR. We then convert both the recognized transcripts and the corresponding reference texts into pinyin using `pypinyin`, and compute CER between the resulting pinyin sequences. The same transcription and pinyin-conversion procedure is used for both models. Higher DNSMOS and UTMOSv2 scores and lower pinyin CER indicate better performance.

Table 4: Automatic speech quality and intelligibility evaluation before and after RL. Pinyin CER is reported as a percentage. Arrows indicate the preferred direction; bold marks the better result for each metric.

Model	DNSMOS ↑	UTMOSv2 ↑	Pinyin CER (%) ↓
Qwen3-TTS-12Hz-1.7B-VoiceDesign	3.40	3.09	1.16
Qwen3-TTS-GRPO	**3.42**	**3.19**	**1.08**

Results. As shown in Table 4, DNSMOS increases from 3.40 to 3.42 and UTMOSv2 from 3.09 to 3.19, while pinyin CER decreases from 1.16% to 1.08%. These modest, consistent improvements across all three automatic metrics suggest that the gains in diversity and instruction following are not accompanied by degraded speech quality or intelligibility in this evaluation. Instead, the results are consistent with small improvements in naturalness, intelligibility, and audio quality, providing additional support for the effectiveness of our RL approach.

Generalization of instruction following. To assess whether the instruction-following gains extend beyond our benchmark, we additionally evaluate the original Qwen3-TTS-12Hz-1.7B-VoiceDesign checkpoint and Qwen3-TTS-GRPO on InstructTTSEval (Huang et al., 2025) [5]. Table 5 reports the instruction following rate (IFR) across the three instruction scales before and after RL.

After training, IFR increases from 83.6% to 86.9% on APS and from 65.0% to 65.4% on RP, corresponding to gains of 3.3 and 0.4 percentage points, respectively, while DSD remains unchanged at 80.6%. These results support the cross-benchmark generalization of the instruction-following improvements induced by our training procedure, rather than gains confined to our own evaluation benchmark. The improvement is most pronounced for APS, with a smaller gain for RP and no degradation on DSD.

[5] Following the approach of the original paper, we use Gemini-2.5-Pro as the LLM judge.

Table 5: Instruction following on InstructTTSEval before and after RL. IFR is reported as a percentage, with higher values indicating better performance. Bold marks the best result in each column, including ties.

Model	APS (%) ↑	DSD (%) ↑	RP (%) ↑
Qwen3-TTS-12Hz-1.7B-VoiceDesign	83.6	**80.6**	65.0
Qwen3-TTS-GRPO	**86.9**	**80.6**	**65.4**

A.4 Ablation Study

We ablate four components of the instruction-gated GRPO procedure: (1) setting the gender gate to $G_i = 1$ for all candidates; (2) setting the instruction-following gate to $F_i = 1$; (3) replacing the diversity bonus with a constant $D_i = 0$; and (4) omitting the APS-style reformulation used for reward evaluation. The first three variants modify Equation 7, retaining the length penalty. In the fourth variant, the reward evaluator uses the original instructions rather than reformulating DSD and RP descriptions into APS-style instructions as described in Section 4.2.

Table 6: Ablation of instruction-gated GRPO on Chinese prompts. IFR is reported as a percentage and VS denotes Vendi Score. Higher is better for both metrics; the best value in each column is bolded. "w/o APS reformulation" removes scoring-time instruction reformulation, not the APS evaluation scale.

	APS		**DSD**		**RP**	
Variant	**IFR (%) ↑**	**VS ↑**	**IFR (%) ↑**	**VS ↑**	**IFR (%) ↑**	**VS ↑**
Full model	83.3	3.456	86.6	4.085	48.4	4.040
w/o gender gate ($G_i = 1$)	81.3	3.663	81.7	4.151	43.2	4.325
w/o Instruction Following gate ($F_i = 1$)	72.4	4.384	77.3	4.676	45.0	4.219
w/o diversity bonus ($D_i = 0$)	83.7	2.467	86.4	2.813	50.1	2.925
w/o APS reformulation	82.1	3.220	83.3	3.712	47.9	3.841

Table 6 shows that setting either the gender gate (G_i) or the instruction-following gate (F_i) to 1 lowers IFR while increasing diversity across all three instruction scales. Disabling either gate relaxes the reward's acceptance criteria, allowing more instruction-violating candidates to receive positive rewards. The resulting diversity gains therefore come at the expense of instruction following. In contrast, fixing the diversity bonus at $D_i = 0$ removes the incentive to generate distinct timbres for the same instruction, leading to a substantial drop in diversity: VS decreases from 3.456 to 2.467 on APS, from 4.085 to 2.813 on DSD, and from 4.040 to 2.925 on RP. Finally, omitting scoring-time APS reformulation reduces IFR across all three scales, consistent with additional errors in instruction-following judgments when the reward evaluator directly assesses the original DSD and RP descriptions. These noisier reward signals weaken the model's instruction-following ability.

A.5 Human Annotation Interfaces

We show the annotation interfaces used for the perceptual validation of within-condition timbre diversity (Section 5.4) and the human evaluation of instruction following (Appendix A.2).

A.5.1 Timbre Diversity Annotation

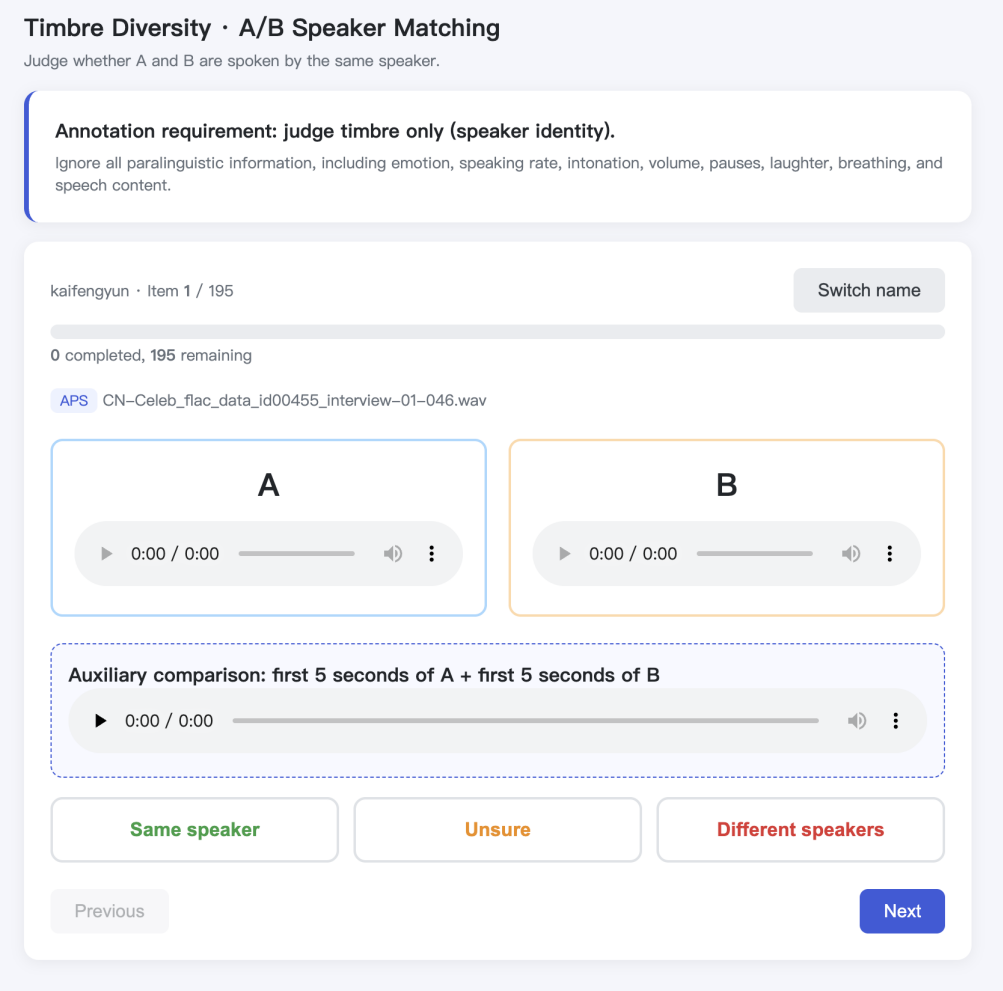

Figure 5: Annotation interface for the timbre diversity experiment. Annotators listen to utterances A and B and select "Same speaker," "Unsure," or "Different speakers." The instructions explicitly ask annotators to judge timbre (speaker identity) only, ignoring other paralinguistic attributes and speech content. An auxiliary player concatenates the first five seconds of each utterance to facilitate direct comparison.

A.5.2 INSTRUCTION-FOLLOWING ANNOTATION

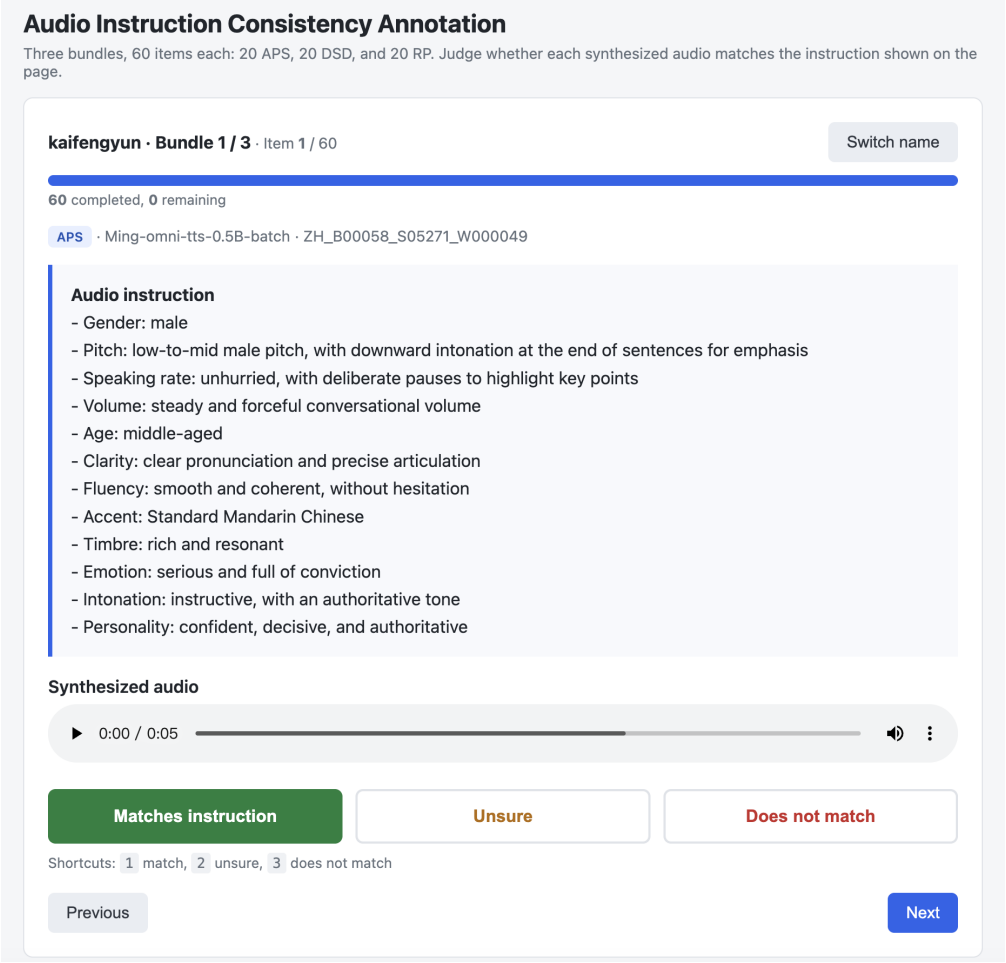

Figure 6: Annotation interface for the instruction-following experiment. Annotators read the displayed audio instruction, listen to the synthesized speech, and select "Matches instruction," "Unsure," or "Does not match." The screenshot shows an APS instruction with attribute-level descriptions of the target voice and speaking style.

A.6 Prompt Templates for Instruction Annotation and Evaluation

We use the speech-captioning, instruction-generation, and LLM-judge templates from InstructTTSEval (Huang et al., 2025). For reproducibility, we reproduce the speech-captioning template from its Figure 3, the DSD template from Figure 4, the two RP templates from Figures 6 and 7, and the English version of the judge template from Figure 8. The prompt wording, examples, and output-format sketches of these templates are retained from the source; line wrapping, typographic symbols, and Markdown code fences are normalized for typesetting. The judge's Markdown table is presented as a list without changing its criteria. We additionally document our APS-reformulation prompt in Appendix A.6.4; it is separate from the InstructTTSEval templates. The placeholder `{style_desc}` denotes the voice-style description provided to the instruction-generation model. Ellipses in the output sketches are part of the published templates, rather than omissions from this appendix.

The shared templates do not imply identical model versions. In our annotation pipeline, Gemini-3.1-Pro-Preview produces the speech captions, and GPT-4o generates the DSD and RP instructions. Gemini-3.1-Pro-Preview also serves as the judge for VoxDICE; the additional InstructTTSEval evaluation in Table 5 uses Gemini-2.5-Pro. The judge template documented here concerns the reported instruction-following evaluation, rather than the separate Qwen3-Omni reward evaluator used during GRPO training.

A.6.1 Speech Captioning and APS Instructions

The speech-captioning prompt describes 12 dimensions: gender, pitch, speed, volume, age, clarity, fluency, accent, texture, emotion, tone, and personality. These explicit descriptions provide the APS instructions and the reference descriptions for DSD and RP generation.

Speech-captioning prompt (InstructTTSEval, Figure 3)

```
You are a Paralinguistics Analysis Expert trained to decode human vocal
  characteristics and speaking styles through acoustic patterns. Here is the
  instruction for your task.

### Instructions for Speech Style Caption

**Objective**: Based on the provided speech, generate a JSON-style output that
  includes concise descriptive phrase of the following vocal characteristcs.
  Please reply in English.

---

### Vocal Characteristics

- Gender: Speech characteristics associated with different gender identities,
  including vocal tract differences and socialized speech patterns.
- Pitch: The perceived frequency of sound, determining whether a voice sounds
  high or low. Typically, male voices have a lower pitch, while female voices
  tend to have a higher pitch. You can express relative pitch levels based on
  gender, for example: ``high female pitch``, ``Low and stable male pitch``.
- Speed: How quickly or slowly someone is speaking, which often changes
  throughout dialogue. If the speaker exhibits specific rhythm patterns, please
  indicate.
- Volume: How loudly or softly someone speaks which can fluctuate significantly.
  Examples range from whispering to normal conversation volume to shouting.
- Age: An inference of the speaker's age group or life stage (such as child,
  teenager, young adult, middle-aged, elderly) based on vocal characteristics.
  If it is challenging to identify the exact stage, you may simply indicate the
  general phase.
- Clarity: Whether pronunciation is distinct and precise or mumbled and blurred.
  Low clarity may involve slurring, mumbling, or running words together, while
  high clarity features precise articulation of sounds.
```

```
- Fluency: The smoothness and continuity of speech, indicating how naturally
  words flow without excessive hesitation, repetition, or filler words (such as
  ``um,`` ``uh,`` ``like,`` ``you know``).
- Accent: Distinctive way of pronouncing words that indicates geographical
  origin, socioeconomic background, or non-native speaker status. If the accent
  is sufficiently pronounced, please specify the dialect region as precisely as
  you can. Otherwise response with the approximate region such as American
  English, British English or Mandarin Chinese.
- Texture: The timbral quality of a speaker's voice, including descriptors such
  as sweet, husky, deep, bright, warm, nasal, mellow, gravelly, or delicate.
  These attributes reflect both physiological traits (e.g., vocal fold
  structure) and stylistic nuances, enabling differentiation between speakers or
  analysis of emotional/expressive tendencies.
- Emotion: The feeling expressed while speaking. which can shift during
  conversation. A person might begin speaking calmly but become increasingly
  frustrated, or switch from sadness to laughter within the same utterance.
- Tone: The emotional or attitudinal quality conveyed through vocal inflection,
  encompassing patterns of pitch variation that signal nuances like sarcasm,
  formality, enthusiasm, or detachment.
- Personality: Please infer the speaker's overall personality based on the
  aforementioned voice characteristics, such as extroversion/introversion,
  confidence, assertiveness, or anxiety. Only describe the speaker's prominent
  and consistent personality traits evident in the speech.

---

### Output Format

{
      ``gender``: ``Gender of the speaker``,
      ``pitch``: ``Description of the speaker's pitch``,
      ``speed``: ``Description of the speaker's temporal flow``,
      ``volume``: ``Description of the speaker's dynamic intensity``
      ......
}

---

### Final Checklist

- Make sure to note how the vocal characteristics change over time throughout
  the speech.
- The description should be concise phrase, avoid using single words or long
  sentences.
- When describing changes in characteristics, you may reference spoken content
  to indicate the timing of the changes, but ensure to summarize the key points
  semantically rather than repeating verbatim.
- When analyzing pitch, speed, and volume, ensure to distinctly differentiate
  their variations rather than uniformly labeling them as ``moderate``. Avoid
  using the word ``moderate``.
- Please utilize diverse vocabulary and varied expressions, avoiding confinement
  to the illustrative examples provided.
```

A.6.2 DSD INSTRUCTION GENERATION

The DSD template requests three stylistically diverse instructions grounded in the supplied voice attributes. Each instruction emphasizes a different combination of four to five attributes.

DSD-generation prompt (InstructTTSEval, Figure 4)

```
Please generate exactly **3 English style instructions** to guide a
  Text-to-Speech (TTS) system strictly based on the following voice style
  description. Return your result in **JSON format** as follows:

```json
{
  ``instructions``: [``Instruction 1...``, ``Instruction 2...``, ``Instruction
  3...``]
}
```

Key Requirements

1. **Fine-grained detail**
  - Each instruction must draw on **4-5 distinct voice attributes** from the
  description (e.g., gender, pitch, rate, volume, clarity, fluency, emotion).
  - If an attribute is especially striking (e.g. ``hoarse,`` ``furious``), give
  it prominence; omit subtler traits if needed.

2. **Maximum diversity**
  - **No two instructions** should begin or end the same way---vary your
  openings, sentence length, word order, and phrasing.
  - Use a rich palette of synonyms and idioms; employ different grammatical
  structures (e.g. active vs. passive, compound vs. simple sentences).
  - Avoid repeating core verbs or adjectives across instructions.
  - **Do not** start any instruction with **Speak**, **Deliver**, **Use**,
  Adopt, **Project**, **Maintain**, **Channel**, **Utilize**.

3. **Idiomatic, fluent English**
  - Do **not** lead with framing words like ``Imagine,`` ``Envision,`` or
  similar.
  - Maintain a **natural monologue style**---not dialogue or questions.

4. **Instruction format**
  - If quoting a phrase from the description (e.g. ``when saying ..., raise your
  voice``), copy it verbatim.
  - Focus each instruction on a **different combination** of attributes---do not
  echo the same group of features.

Voice Style Reference

```json
{style_desc}
```

Please return **only** the JSON object---no extra text.
```

A.6.3 RP INSTRUCTION GENERATION

We use two RP templates. The first asks for character-centered director's notes grounded in the voice description; the second requests a minimal character description paired with an action or speaking manner. Both templates request an explicit inference of plausible personas before the instruction list.

RP-generation prompt I (InstructTTSEval, Figure 6)

```
Please generate a total of **3 natural English role-play instructions** to guide
  a Text-to-Speech (TTS) system strictly based on the following voice style
  description. Do not introduce any traits not present in the description or add
  creative embellishments.

---
### Output Format Requirements
First, analyze the voice style description and briefly summarize your
  understanding and reasoning about the possible character personas.
Then output **only** the JSON object in this exact structure:
```json
{
  ``instructions``: [``Instruction 1...``, ``Instruction 2...``, ``Instruction
  3...``]
}
```

Instruction Guidelines

1. Each instruction must embody a **specific character persona** inferred
  directly from the voice style description.
2. Write each as a **natural, colloquial English phrase** describing the
  character and their manner of speaking---do not provide actual dialogue lines.
3. Phrase instructions like a director's note to an actor: concise, vivid, and
  instantly actionable.
4. Vary sentence structures and points of entry; avoid formulaic templates.
5. Use declarative sentences only; do **not** start with framing words such as
  ``like,`` ``imagine,`` ``pretend,`` ``as,`` ``you are,`` ``acting,`` etc.

Good Instruction Examples

* ``Step into the shoes of a big sister at a family gathering, balancing
  patience and firmness as you restore order.``
* ``Portray a determined professional in a boardroom, your voice steady with
  just a hint of nervous tension.``
* ``Embody the excitement of an adventurer discovering a hidden treasure,
  speaking with wide-eyed enthusiasm.``

Bad Instruction Examples (avoid)

* [x] ``Pretend you are a cold assassin.``
* [x] ``Act like a weary teacher.``
* [x] ``Speak in a calm voice with moderate pace.``

Voice Style Reference Description

```json
{style_desc}
```
```

RP-generation prompt II (InstructTTSEval, Figure 7)

```
Please generate **3 minimal, natural English role-play instructions** based on
  the following voice style description to guide a Text-to-Speech (TTS) system
  in synthesizing the corresponding character style voice.

---
## Output Format
First, outline your reasoning: explain how you inferred possible character
  personas from the voice style description.
Then output **only** this JSON object:
```json
{
  ``instructions``: [``Instruction 1``, ``Instruction 2``, ``Instruction 3``]
}
```

Generation Rules

1. Begin by deducing which character personas fit the voice style description.
  You may choose imaginative roles (e.g., ``general,`` ``empress,`` ``knight,``
  ``poet,`` ``kitten,`` ``robot``), but they must align with the description.
2. Each instruction must consist of **``Character + a single minimal action or
  speaking manner``** and must not include any acoustic details.

  * [OK] Example: ``On the debate podium, the lawyer unleashes a passionate
  argument.``
  * [OK] Example: ``With swift, clear enunciation, the seasoned commentator
  brings the match to life.``
  * [OK] Example: ``A rural poet slowly recounting the town's legend.``
3. Use colloquial, vivid phrasing---like a director's off-the-cuff cue.
4. Vary sentence structures; avoid templated patterns. All three instructions
  must differ and must not start the same way.
5. Write each as one concise English sentence.

Voice Style Reference Description:

```json
{style_desc}
```
```

A.6.4 APS REFORMULATION FOR REWARD EVALUATION

For the scoring-time reformulation described in Section 4.2, DSD and RP instructions are converted into APS-style descriptions without changing the policy input. Only requirements conveyed by the corresponding instruction are retained; the original APS provides concrete wording rather than additional constraints. The 12 canonical APS names are `gender`, `pitch`, `speed`, `volume`, `age`, `clarity`, `fluency`, `accent`, `texture`, `emotion`, `tone`, and `personality`. The following author-provided prompt specifies the JSONL conversion and resumable processing. In its JSON field example, `\n` denotes an escaped newline within a string; colons and surrounding whitespace have been normalized for presentation.

APS-reformulation prompt (DSD/RP to APS)

```
Convert DSD and RP prompts into APS-style instructions for audio judging.

The input and output are JSONL files. Each output row preserves the input and
adds:

    ``judge_aps``: {
        ``DSD``: ``gender: male\n\nspeed: relatively slow\n\ntone: like a
  command``,
        ``RP``: ``emotion: firm\n\ntone: undeniable``
    }

Each generated line must use one of the 12 canonical APS names before the
colon. Only requirements conveyed by the corresponding DSD/RP instruction are
retained, while the original APS supplies concrete, judge-friendly wording.
The output file and an append-only cache are reusable, allowing interrupted
runs to resume without relabeling successful cases.
```

A.6.5 LLM-JUDGE PROMPT

The same judge template evaluates instruction–audio consistency across APS, DSD, and RP. It first elicits descriptions of 12 vocal dimensions and then a single binary `Consistency` judgment, rather than separate pass/fail scores for each dimension. In the judge template, *Speech Rate*, *Timbre Quality*, and *Intonation* correspond to *Speed*, *Texture*, and *Tone* in the captioning template. Attributes not specified by an instruction impose no constraints. IFR is the fraction of evaluated instruction–audio pairs whose `Consistency` judgment is `True`.

LLM-judge prompt (InstructTTSEval, Figure 8; English version)

```
You are an expert with rich acoustic knowledge. Please describe a speech segment
  according to the following dimensions and judge whether the speech matches the
  given description, outputting **True** (matches) or **False** (does not match)
  on the consistency dimension, ignoring non-style factors (sound quality,
  naturalness, etc.).

---

## Evaluation Dimensions
- Gender: Speech characteristics related to different gender identities,
  including vocal cord differences and socialized speech patterns.
- Pitch: The perceived frequency of the voice, determining whether the voice is
  high or low. Typically, male voices have lower pitch, female voices have
  higher pitch. Relative pitch levels can be based on gender expression, e.g.,
  ``female high voice,`` ``male deep stable voice.``
- Speech Rate: The speed of talking, which often varies in conversation. If the
  speaker exhibits specific rhythmic patterns, please indicate.
- Volume: The loudness or softness of speech, which can vary greatly. Examples
  include whispering, normal conversational volume, or shouting.
- Age: Inferring the speaker's age group or life stage (such as child, teenager,
  young adult, middle-aged, elderly) based on speech characteristics. If
  difficult to determine a specific stage, simply indicate the approximate stage.
- Clarity: Whether pronunciation is clear and accurate or unclear. Low clarity
  may involve murmuring, mumbling, or connected speech, while high clarity is
  characterized by precise pronunciation.
- Fluency: The fluency and continuity of language, reflecting whether words flow
  naturally without excessive hesitation, repetition, or filler words (such as
  ``um,`` ``ah,`` ``like,`` ``you know``).
- Accent: The unique way of pronunciation, reflecting geographical origin,
  socioeconomic background, or non-native speaker identity. If the accent is
  distinctive enough, please specify the dialect region as precisely as
  possible; otherwise, indicate the general region, such as American English,
  British English, or Standard Mandarin.
- Timbre Quality: The timbral quality of the voice, including descriptions like
  sweet, hoarse, deep, bright, warm, nasal, soft, rough, or thin. These
  attributes reflect physiological characteristics (such as vocal cord
  structure) and stylistic nuances, useful for distinguishing speakers or
  analyzing emotional/expressive tendencies.
- Emotion: The emotion expressed when speaking, which may change during the
  conversation. For example, a person might start speaking calmly but gradually
  become upset, or transition from sadness to laughter within the same sentence.
- Intonation: The emotional or attitudinal quality conveyed through voice
  modulation, including pitch variation patterns, expressing nuances such as
  sarcasm, formality, enthusiasm, or indifference.
- Personality: Inferring the overall personality of the speaker based on the
  above speech characteristics, such as extroverted/introverted, confident,
  assertive, or anxious. Only describe personality traits that are obvious and
  consistent in the speech.

---

## Judgment Criteria
```

```
**True (Matches)**
The speech sample satisfies the main stylistic features of the description:
- Major style dimensions (such as gender, pitch, speech rate, emotion, etc.) are
  consistent with the description
- No obvious deviations or conflicts

**False (Does not match)**
The speech sample fails to satisfy the main stylistic features of the
  description:
- There is at least one key stylistic feature that obviously conflicts with the
  description
- The overall listening impression deviates from the described style

---

## Notes
* When there is an obvious and objective mismatch in a dimension, such as when
  the speaker's gender or age conflicts with the description, directly judge it
  as False
* The description is very likely to have obvious conflicts, degree
  inconsistencies, or complete mismatches with the speech; do not readily trust
  the given description, you should first retain your own understanding of the
  speech
* The speaker's gender is especially likely to be contrary to the description,
  please pay special attention
* When terms indicating degree such as ``excited,`` ``intense,`` etc. are
  mentioned in the description, extra attention is needed, as the specified
  dimension in the speech may not be as intense as in the description, such as
  emotions not being excited enough, pitch not being high enough, volume not
  being loud enough, etc., in which case it should be judged as False
* Only evaluate **style consistency**, ignore non-style factors such as
  pronunciation accuracy, naturalness, etc.
* Use the description as the sole basis, without personal subjective preferences
* For characteristics not mentioned in the description, there are no
  restrictions on that feature, and it should not affect the judgment
* When the description focuses on only one dimension (such as emotion), judgment
  should focus on that dimension

---

## Output Format Requirements:
Please strictly use JSON format, containing a dictionary with the following
  structure:
```json
{
  ``Gender``: ...,
  ``Pitch``: ...,
  ``Speech Rate``: ...,
  ...
  ``Consistency``: True/False
}
```
```