UniSRM: A Unified Speech Reward Model for Reasoning-Based Fine-grained Assessment

Yuanyuan Wang[1], Dongchao Yang[1], Yayue Deng[1],
Zhiyong Wu[1,2,†], Yiwen Guo[3], Helen Meng[1], Xixin Wu[1,†],

[1] The Chinese University of Hong Kong, [2] Tsinghua University, [3] Independent Researcher

Abstract

Evaluating speech generation still relies heavily on human judgments, such as Mean Opinion Score (MOS), which are expensive, subjective, and difficult to reproduce at scale. While a few recent studies have begun to explore AudioLLM-based judge models, existing efforts typically target only a narrow set of scenarios (e.g., utterance-level quality or single-turn dialogue) and provide limited coverage of diverse speech generation tasks and evaluation dimensions. In this work, we propose UniSRM, a unified speech reward model that can support multi-dimensional, interpretable reward signals with reliable reasoning. To support training and evaluation, we introduce UniSRM-Data and UniSRM-Bench, covering speech evaluation tasks from utterance-level quality to context-level coherence. Based on this dataset, we present the unified speech reward model, UniSRM, with a two-stage pipeline that enables reasoning-based fine-grained assessment. Furthermore, we introduce Reasoning-Consistent Rewards to improve the reliability of the reasoning process. Experiments show that UniSRM delivers more reliable and human-aligned judgments across a broad range of speech evaluation tasks, offering a practical foundation for scalable and unified evaluation of speech quality [1].

1 Introduction

Autoregressive large language models (LLMs) have revolutionized natural language processing with their strong generation and reasoning capabilities (Brown et al., 2020; Touvron et al., 2023; Team et al., 2023). Reinforcement learning (RL) (Ouyang et al., 2022; Schulman et al., 2017) has emerged as a promising framework for improving model alignment with human preferences through reward-based feedback. However, in speech generation, the lack of well-designed reward models makes optimization particularly challenging. The Mean Opinion Score (MOS) (ITU-T, 1996) is an ideal, widely accepted criterion for assessing speech quality. However, it is costly to collect and inherently subjective, with ratings varying across listeners and datasets, which lack reliable optimization targets and hinder reproducible comparisons.

Existing speech generation methods powered by RL mainly rely on two categories of reward signals. First, some works (Zhang et al., 2024; Fu et al., 2024; Hu et al., 2024; Sun et al., 2025; Chen et al., 2025b) utilize classical speech reward signals, which are typically objective metrics such as WER, speaker similarity (SIM), and UTMOS(Saeki et al., 2022). While these metrics provide effective indicators, each of them captures only a *single aspect* of speech. For instance, WER mainly reflects textual correctness, and SIM measures timbre similarity to a reference speaker. As a result, such rewards cannot holistically evaluate speech. Moreover, when used as reward signals, they are often treated as black-box scorers that output a single scalar value without intermediate and explicit explanations. This *lack of transparency* may introduce bias and inconsistency into the optimization process, ultimately undermining the reliability of reward-based alignment for speech generation. Secondly, some works use Large Audio Language Models (LALMs) as speech judges. For example, WavReward (Ji et al., 2025) and SageLM (Ge et al., 2025) fine-tune LALMs for single-turn spoken dialogue evaluation. SpeechJudge (Zhang et al., 2025) trains a generative reward model largely centered on naturalness over utterance-level speech. These works remain *limited in task coverage*, typically focusing on only utterance-level speech or single-turn dialogue. They also have incomplete evaluation dimensions, such as overlooking speaker similarity. Moreover, the rule-based RL (Ge et al., 2025) provides limited supervision over reasoning, which

† Corresponding authors.

[1] The checkpoint and dataset are publicly available at https://github.com/lavendery/UniSRM.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46346–46366
July 2-7, 2026 ©2026 Association for Computational Linguistics

cause inconsistency between the generated rationales and final decisions. Overall, current research on speech reward modeling still faces four main challenges: (1) lack of transparency in scoring; (2) incomplete evaluation dimensions; (3) limited task coverage; and (4) insufficient supervision over the reasoning process.

To address these limitations, we propose UniSRM, a unified speech reward model designed to produce multi-dimensional, interpretable reward signals backed by reliable reasoning. Experiments demonstrate that UniSRM yields judgments that are not only more reliable but also better aligned with human preferences. Our main contributions are summarized as follows:

- **Comprehensive Data and Benchmark.** We construct UNISRM-DATA, a unified dataset that covers speech evaluation tasks *from utterance-level quality to context-level coherency*. In parallel, we introduce UNISRM-BENCH as a comprehensive benchmark for unified speech reward modeling.
- **A Unified Reward Model with Explicit Decomposition.** Building on UNISRM-DATA, we develop UNISRM with a two-stage training pipeline, which explicitly decomposes speech assessment into multiple dimensions, enabling fine-grained evaluation across diverse tasks.
- **Reasoning-Consistent RL Optimization.** During the RL stage, we propose RCR-GRPO (Reasoning-Consistent Rewards), which assigns rewards at the dimension-wise reasoning process to improve reliability.

2 Related Work

2.1 Multimodal Reward Models

Reinforcement learning from human feedback (RLHF) and its variants have become an effective paradigm for aligning large language models (LLMs) with human preferences (Ziegler et al., 2019; Ouyang et al., 2022; Rafailov et al., 2023). Recent works extend this paradigm to multimodal settings, where reward models provide supervisory signals for both understanding and generation tasks over image, video, and audio (Team et al., 2023; Lee et al., 2023; Yang et al., 2023; Wang et al., 2024a,b; Yang et al., 2024b, 2025a; Liu et al., 2025; Zhao et al., 2025; Wang et al., 2025d,e,f; Yang et al., 2025c, 2026). However, despite this progress in multimodal settings, reward modeling for speech still has meaningful room for further improvement (Yang et al., 2024a, 2025b).

2.2 Speech Reward Models

Using Large Audio Language Models (LALMs) as automated speech judges has recently received growing attention. ATT (Wang et al., 2025c) and ALLD (Chen et al., 2025a) both introduce a human-likeness speech evaluation corpus and train LALMs to describe and score speech quality in a human-aligned manner. QualiSpeech (Wang et al., 2025b) develops a detailed dataset for low-level speech quality assessment. AudioJudge (Manakul et al., 2025) explores prompting strategies to elicit multi-aspect judgments. WavReward (Ji et al., 2025) extends LALMs to evaluate both IQ and EQ for spoken dialogue systems, but is restricted to single-turn dialogue. SageLM (Ge et al., 2025) also trains an end-to-end spoken dialogue evaluator via SFT for single-turn conversational quality. SpeechLLM-as-Judges (Wang et al., 2025a) fine-tunes a speech quality LLM on a large-scale SpeechEval dataset to perform assessment, comparison, improvement suggestion, and deepfake detection. SpeechJudge (Zhang et al., 2025) trains a generative reward model, aiming at utterance-level preference evaluation over paired speech samples.

Some of these above approaches offer evaluation dimensions that are not sufficiently fine-grained and comprehensive (Wang et al., 2025c; Chen et al., 2025a; Wang et al., 2025b), while others inherently inherit the limited understanding capacity of the underlying LAMs and may therefore produce shallow judgments (Manakul et al., 2025), and some are restricted by insufficient coverage of task scenarios (Wang et al., 2025a, 2026; Zhang et al., 2025; Ge et al., 2025; Ji et al., 2025). Moreover, rule-based reinforcement learning lacks supervision over the reasoning process, which can lead to inconsistency between the rationale and final result (Ge et al., 2025; Zhang et al., 2025).

In this paper, we propose an end-to-end speech reward model, UniSRM, that decomposes speech quality into multiple complementary dimensions and generates explicit reasoning traces before producing an aggregated preference decision, thus offering both a richer supervision signal and improving interpretability.

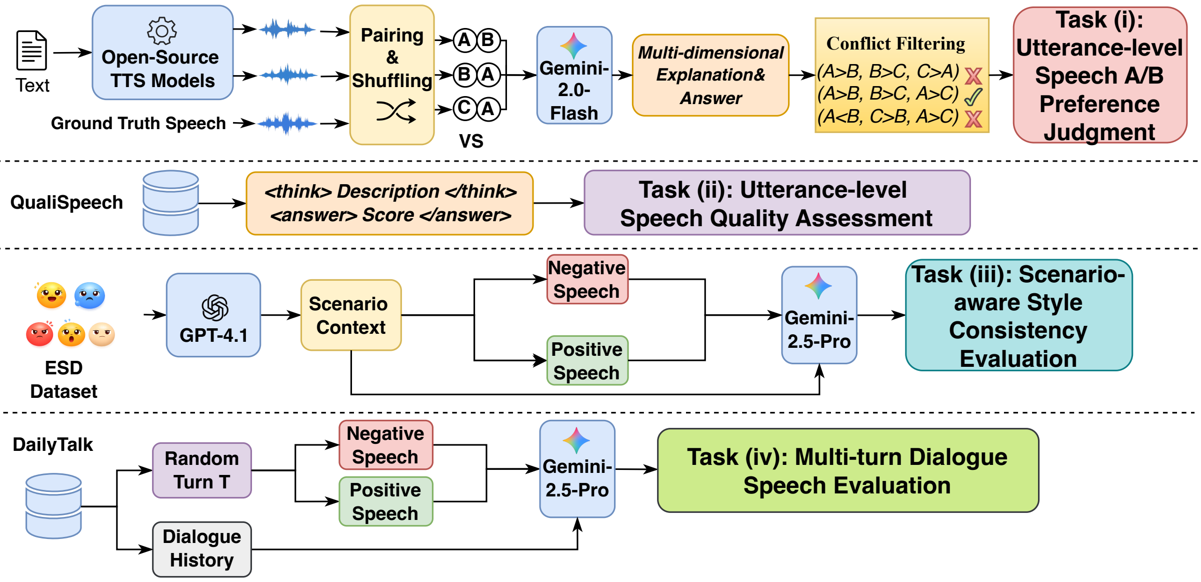

Figure 1: The Pipeline of UniSRM Dataset construction.

3 UniSRM-Data and UniSRM-Bench

As shown in Figure 1, we construct a unified dataset that covers speech evaluation tasks ranging *from utterance-level quality to context-level coherency*, including: (i) utterance-level speech A/B preference judgment, (ii) utterance-level speech quality assessment, (iii) scenario-aware style coherency evaluation conditioned on textual context, and (iv) multi-turn dialogue speech evaluation conditioned on dialogue history. In the following, we describe the construction process of each task in detail.

Task 1: Utterance-Level Speech A/B Preference Judgment. Pairwise preference comparison is often easier and more reliable than absolute scoring for speech quality assessment, as it directly asks which of two speech samples of the same target text is better overall. Such relative ranking, compared to absolute scores, is less susceptible to variations between judges.

Given a text prompt x, we use multiple open-source TTS models s to synthesize diverse candidate speech signals.

$$\mathcal{S}(x) = \{s_1, s_2, \ldots, s_K\}, \tag{1}$$

where K is the number of open-source TTS models. These speech samples, generated from the same textual content by different synthesis models, are paired to form comparison candidates. We additionally include the ground-truth recording S^{GT} as another candidate. We then form unordered comparison pairs:

$$\mathcal{P}(x) = \Big\{(s_A, s_B) \;\Big|\; s_A, s_B \in \mathcal{S}(x) \cup \{S^{\mathrm{GT}}\}, \\ s_A \neq s_B\Big\}. \tag{2}$$

To obtain preference annotations, we employ Gemini-2.0-Flash to generate multi-dimensional scores and explanations for each candidate in a pair, covering *Text Fidelity & Intelligibility, Speaker Similarity, Prosody & Expressiveness, and Naturalness & Audio Quality*, followed by a final binary decision indicating which speech is better. For each speech sample, Gemini assigns a score in $[0, 10]$ for each dimension, and the total score is computed by summation. The final preference label is determined by comparing the total scores:

$$\ell_{better} = \begin{cases} speechA, & \text{if } T_A > T_B, \\ speechB, & \text{otherwise.} \end{cases} \tag{3}$$

This process yields a high-quality dataset of speech preference pairs for speech reward modeling. In *Appendix* G, we present more details about how to remove nosiy samples and human verification criteria to improve the quality of dataset.

Task 2: Utterance-Level Speech Quality Assessment. For single-sample speech quality assessment, we directly leverage the public QualiSpeech (Wang et al., 2025b) dataset, which provides MOS-like annotations across seven perceptual aspects: *Noise, Distortion, Speed, Continuity, Naturalness, Listening effort, and Overall quality* (each in $[1, 5]$). To align with our reasoning-based

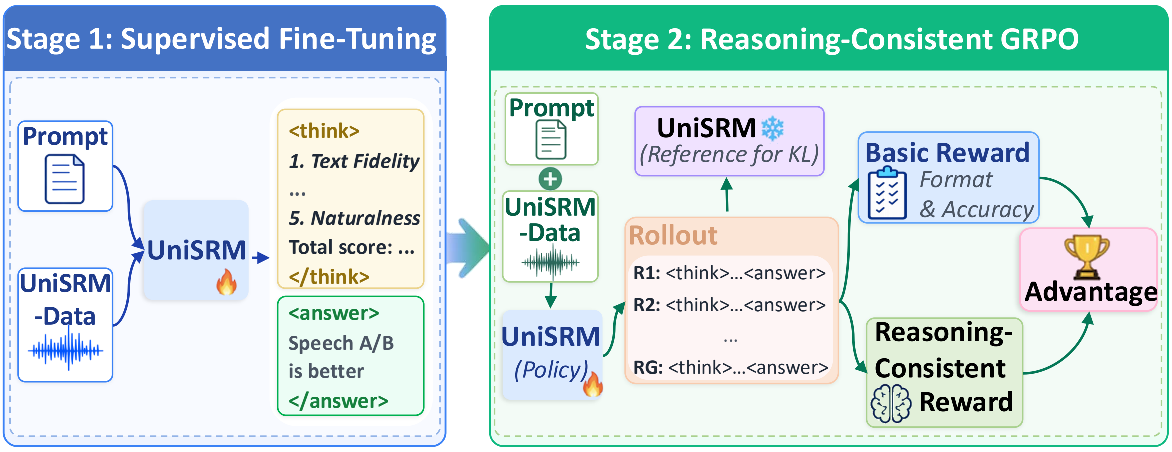

Figure 2: Our proposed two-stage framework of UniSRM.

training format, we treat QualiSpeech's textual rationales as the model's intermediate reasoning. Specifically, *Noise description, Distortion description, Unnatural pause, Feeling of voice, together with a Natural language description* are used as the reasoning trace, while the seven aspect scores are used as the final structured answer. This design enables UniSRM to learn fine-grained, interpretable quality assessment rather than a black-box scalar prediction.

Task 3: Scenario-Aware Style Coherency conditioned on Textual Context. Speech quality cannot always be judged from an isolated utterance: in expressive narration or voice acting, speech should match the scene context and target emotion. We therefore define scenario-aware style consistency evaluation, where a model compares two speech samples conditioned on a textual scenario and paragraph context.

Using ESD (Zhou et al., 2022), we treat each original recording as the positive sample, and use GPT-4.1 (OpenAI et al., 2024) to generate a coherent scenario description and paragraph context conditioned on the utterance text and emotion label. Hard negatives are created by sampling one corruption type per instance: (i) real-speech negatives from ESD with mismatched text and/or emotion, or (ii) TTS negatives synthesized by open-source TTS systems, optionally with GPT-4.1–generated mismatching text under the same scenario.

For TTS negatives, we assign prompt speech from ESD or LibriTTS-R (Koizumi et al., 2023), enforce emotion mismatch, and randomize controllable attributes (e.g., role, speaking rate, style) while filtering accidental emotion matches. Finally, given the scenario context, we use Gemini-2.5-Pro to provide multi-dimensional scores and rationales along *Text Fidelity, Scenario Style Match, Naturalness/Audio Quality*. In this way, we generate a bilingual dataset for both Chinese and English.

Task 4: Multi-Turn Dialogue Speech Evaluation conditioned on Dialogue History. Existing dialogue speech reward modeling is often restricted to single-turn settings. We introduce a multi-turn dialogue speech evaluation task conditioned on spoken dialogue-history, enabling reward modeling of dialogue-level consistency.

Based on DailyTalk (Lee et al., 2022), for each dialogue, we randomly sample a target turn t and treat the speech of turns $1, \ldots, t-1$ as the dialogue history. The original speech at turn t is used as a positive candidate. To reduce shortcut learning from recording conditions (e.g., *real vs. synthesized*), we also synthesize a subset of positives using the open-source TTS systems as negatives.

We construct hard negatives along two axes: text mismatch and audio mismatch. We randomly sample from three categories: (i) text-only negatives generated by GPT-4.1 conditioned on the multi-turn history (e.g., intent/consistency errors), (ii) audio-only negatives that keep the text fixed but alter speaker/prosody/emotion (e.g., speaker swap or prompt speaker mismatch), and (iii) mixed negatives that combine both. All candidates are synthesized with multiple open-source TTS systems to increase diversity. Given the dialogue-history and two candidate samples, we use Gemini-2.5-Pro to provide multi-dimensional scores and rationales along *Intent Matching, Speaker Consistency, Contextual Consistency, Emotion, Naturalness*.

In conclusion, we partition UNISRM-DATA into three disjoint subsets: supervised fine-tuning $\mathcal{D}_{\text{SFT}}$, RL training $\mathcal{D}_{\text{GRPO}}$, and a test set UNISRM-BENCH. To ensure high label reliability for both optimization and evaluation, we additionally perform human verification on *both* $\mathcal{D}_{\text{RL}}$ and UNISRM-BENCH, retaining only samples that match the majority-vote human preference. Detailed human verification criteria is provided in Appendix G.2. We summarize the task-specific evaluation dimensions in Appendix B (Table 8).

4 Method

In this section, we present the complete training pipeline of UniSRM, as illustrated in Figure 2. Initially, we utilize SFT to adapt the model to diverse evaluation tasks and standardize the output scoring format. Subsequently, to further align the model with human preferences and encourage diversity in the reasoning process, we employ RL on a manually curated high-quality dataset $\mathcal{D}_{\text{RL}}$.

4.1 Supervised Fine-Tuning (SFT)

Speech understanding models inherently can evaluate generated speech from multiple perspectives with rational and interpretable judgments, making them well-suited for speech reward modeling. Therefore, we adopt Qwen2.5-Omni-7B-thinker (Xu et al., 2025) as the backbone and modify its system prompts (Appendix A) to enforce a deterministic and structured output format. As shown in Stage 1 of Figure 2, we train a multi-task speech reward model that supports multi-dimensional reasoning based on $\mathcal{D}_{\text{SFT}}$. This design makes the model produce *dimension-wise evidence* before outputting the final preference/score, which is critical for interpretability and provides a stable foundation for subsequent RL optimization.

Although our data cover four tasks with different inputs, they can be unified as a conditional generation problem. Each training instance is represented as (x, o), where x is the task-specific input prompt, e.g., text content and audio clips, and o is the target structured output. Given an input x, UniSRM outputs a two-part response:

$$\begin{aligned} o &= \pi_\theta(x) \\ &= \texttt{<think>}\,\hat{r}\,\texttt{</think><answer>}\,\hat{y}\,\texttt{</answer>}. \end{aligned} \tag{4}$$

Here, $\hat{r}$ is an explicit reasoning trace containing *dimension-wise* scores and short explanations, e.g., text fidelity, speaker similarity and so on. The final output $\hat{y}$ is task-dependent: it is a binary preference decision for pairwise tasks (Task 1/3/4), or a MOS-like structured score for the pointwise quality task (Task 2). We fine-tune the model using standard autoregressive maximum likelihood on the full target sequence o:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,o)\sim\mathcal{D}_{\text{SFT}}} \sum_{t=1}^{|o|} \log \pi_\theta(o_t \mid o_{<t}, x). \tag{5}$$

After SFT, the resulting policy π_θ serves as a stable initialization for the GRPO stage, where we further improve the correctness, diversity and reliability of the reasoning process.

4.2 Reinforcement Learning with GRPO

While SFT teaches UniSRM to imitate the judge-generated rationales and final decisions, it does not explicitly optimize reward-aligned correctness, and may lead the model to learn fixed pattern reasoning. Therefore, we further optimize UniSRM with Group Relative Policy Optimization (GRPO) to improve reasoning diversity and reliability.

As shown in Figure 2, for each training prompt x, we sample G responses from the current policy π_θ:

$$o^{(g)} \sim \pi_\theta(\cdot \mid x), \quad g = 1, \dots, G, \tag{6}$$

where each output follows our structured format (Eq. 4). We combine three complementary reward components:

$$R(x, o) = \lambda_{\text{fmt}} R_{\text{fmt}}(o) + \lambda_{\text{acc}} R_{\text{acc}}(o) + \lambda_{\text{rc}} R_{\text{rc}}(o), \tag{7}$$

where R_{fmt} indicates format reward, R_{acc} optimizes the final answer accuracy, and R_{rc} provides reasoning-consistent rewards over each dimension. For the format reward R_{fmt}, we define $R_{\text{fmt}}(o) \in \{-1, 0\}$ as an indicator of whether o matches the output format of different tasks in Appendix A. If the output violates the required format or fails parsing, we assign a negative reward -1 to penalize invalid reasoning traces. For the accuracy reward R_{acc} of pairwise tasks (Task 1/3/4), we set

$$R_{\text{acc}}(o) = \mathbf{1}\left[y^{(g)} = y^\star\right], \tag{8}$$

where $y^\star \in \{\text{A}, \text{B}\}$ is the ground-truth preference label. For R_{acc} of Task 2, we use a normalized distance reward on the *overall* score:

$$R_{\text{acc}}(o) = 1 - \frac{|\hat{m}_{\text{overall}} - m^\star_{\text{overall}}|}{m_{\max} - m_{\min}}, \tag{9}$$

and clamp it to $[0, 1]$, where $(m_{\min}, m_{\max}) = (1, 5)$ by default, $\hat{m}_{\text{overall}}$ and $m^\star_{\text{overall}}$ denote predicted and ground-truth overall score.

Model	T1 *acc* ↑	T2 *acc* ↑ /*pcc* ↑	T3-En *acc* ↑	T3-Zh *acc* ↑	T4 *acc* ↑
Objective Metrics					
WER	59.24	-/-	61.44	56.92	84.10
SIM	47.99	-/-	-	-	-
UTMOS	50.20	-/0.449	33.21	48.19	40.48
DNSMOS	49.80	-/0.274	53.51	63.04	50.79
Proprietary Models					
GPT-4o-Audio (Hurst et al., 2024)	61.04	24.60/0.060	64.02	64.82	71.96
Gemini-2.5-Flash	60.44	34.50/0.522	65.68	71.74	71.43
Gemini-2.5-Pro	60.67	28.93/0.517	67.31	63.47	82.40
Open-Source Models					
Kimi-Audio-7B (Ding et al., 2025)	52.81	22.93/0.209	71.22	69.70	64.29
MiMo-Audio-7B (Xiaomi, 2025)	50.40	26.36/0.158	47.97	42.49	59.52
Qwen2.5-Omni-7B (Xu et al., 2025)	51.20	24.03/0.289	49.45	52.17	56.35
SpeechJudge (Zhang et al., 2025)	57.20	-/-	-	-	-
Proposed Method					
UniSRM(Ours)	65.06	39.74/0.551	85.61	91.30	88.89

Table 1: Overall results on UNISRM-BENCH. **T1**: utterance-level pairwise preference judgement. **T2**: fine-grained speech quality scoring. **T3-En / T3-Zh**: scenario-aware style consistency preference in English and Chinese. **T4**: multi-turn dialogue speech evaluation conditioned on spoken dialogue history.

Model	T1	T2	T3-En	T3-Zh	T4
UniSRM(Ours)	65.06	39.74	85.61	91.30	88.89
w/o RCR-GRPO	60.44	37.58	80.81	81.42	82.54
w/o GRPO	60.24	39.20	67.16	70.95	74.60

Table 2: Ablation results over all tasks. Results are reported as accuracy (↑, %).

Reasoning-Consistent Rewards (RCR-GRPO). A key challenge of RL on judge-style rationales is that optimizing only the final answer may encourage shallow or inconsistent reasoning, e.g., a correct label but mismatched reasoning. To address this, we introduce Reasoning-Consistent Rewards R_{rc} to directly supervise the dimension-wise scoring behavior inside <think>.

For pairwise tasks (Task 1/3/4), each output contains dimension-wise scores for both candidates, e.g., $\mathbf{a} = [a_1, \dots, a_D]$ for Speech A and $\mathbf{b} = [b_1, \dots, b_D]$ for Speech B. $D = 4, 3, 5$ is the number of dimensions for task 1/3/4, respectively. We compute a *dimension-wise preference consistency* reward:

$$R_{\text{rc}}(o) = \frac{1}{D}\sum_{i=1}^{D} \mathbf{1}\Big[\text{sign}(a_i - b_i) = \text{sign}(a_i^\star - b_i^\star)\Big]. \tag{10}$$

where $(\mathbf{a}^\star, \mathbf{b}^\star)$ are the ground-truth dimension scores, and $\text{sign}(\cdot) \in \{-1, 0, +1\}$. Intuitively, R_{rc} encourages UniSRM to produce aligned per-dimension comparisons rather than only matching the final preference label, which can improve reasoning reliability.

For speech quality assessment (Task 2), the output <answer> provides a $D = 7$ aspect score vector $\hat{\mathbf{m}} \in \{1, \dots, 5\}^D$. We compute a normalized reward:

$$R_{\text{rc}}(o) = 1 - \frac{1}{D}\sum_{k=1}^{D} \frac{|\hat{m}_k - m_k^\star|}{m_{\max} - m_{\min}}, \tag{11}$$

and clamp it to $[0, 1]$. This reasoning-consistent rewards provide targeted supervision over the intermediate reasoning process, improving diversity while reducing inconsistencies between generated rationales and final decisions.

Group-wise advantage normalization. As shown in Equation 12, we compute advantages by normalizing rewards within each prompt group, where $\mu(x)$ and $\sigma(x)$ denote the mean and standard deviation of $\{R^{(g)}\}_{g=1}^{G}$. This relative normalization reduces reward scale sensitivity and encourages meaningful comparisons across rollouts.

$$A^{(g)} = \frac{R^{(g)} - \mu(x)}{\sigma(x) + \epsilon}. \tag{12}$$

Model	Text	Sim	Expressiveness	Naturalness	AVG
UniSRM(Ours)	83.33	62.25	61.24	43.98	62.70
w/o RCR-GRPO	76.89	59.22	60.23	39.76	59.03
w/o GRPO	83.53	57.83	59.84	42.37	60.89

Table 3: Multi-dimensional results of pair-wise speech preference task. Results are reported as accuracy (↑, %).

Model	Noise	Distortion	Speed	Continuity	Effort	Naturalness	Overall	AVG
QualiSpeech	0.686	0.518	<u>0.250</u>	0.459	0.475	<u>0.486</u>	**0.572**	0.492
UniSRM								
UniSRM(Ours)	**0.754**	**0.547**	0.209	**0.526**	<u>0.478</u>	0.473	<u>0.551</u>	**0.505**
w/o RCR-GRPO	0.688	<u>0.528</u>	0.233	<u>0.512</u>	0.446	0.418	0.542	0.481
w/o GRPO	<u>0.714</u>	0.514	**0.268**	0.471	**0.481**	**0.506**	0.534	<u>0.498</u>

Table 4: Multi-dimensional results on the QualiSpeech dataset. Results are reported as PCC(↑).

GRPO objective with KL regularization. Let $\pi_{\theta_{\text{old}}}$ denote the policy used to generate rollouts. GRPO optimizes a clipped policy gradient objective:

$$\mathcal{J}(\theta) = \mathbb{E}_x \, \mathbb{E}_{g=1}^{G} \Big[\min\Big(\rho_\theta^{(g)} A^{(g)}, \text{clip}(\rho_\theta^{(g)}, 1-\epsilon, 1+\epsilon) \, A^{(g)} \Big) \Big]. \tag{13}$$

where $\rho_\theta^{(g)} = \frac{\pi_\theta(o^{(g)}|x)}{\pi_{\theta_{\text{old}}}(o^{(g)}|x)}$. To prevent excessive drift from the supervised initialization, we add a KL penalty against a reference policy π_{ref} (the SFT model checkpoint):

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathcal{J}(\theta) + \beta \cdot \mathbb{E}_x \, \mathbb{E}_{g=1}^{G} \Big[\text{KL}\Big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x) \Big) \Big]. \tag{14}$$

5 Experiments and Analyses

5.1 Evaluation Metrics

We evaluate UniSRM using two primary metrics: Accuracy (ACC) for pairwise decision tasks and Pearson Correlation Coefficient (PCC) for score-based assessment. For the preference-based tasks (Task 1, Task 3, and Task 4), the model outputs a binary decision $\hat{y} \in \{\text{Speech A}, \text{Speech B}\}$ in the <answer> tag. We report ACC as the proportion of samples whose predicted preference matches the ground-truth label. For the fine-grained speech quality task (Task 2), the model predicts MOS-like scores for multiple aspects. Following the standard practice in QualiSpeech(Wang et al., 2025b), we use PCC to measure how well the predicted scores correlate with the ground-truth scores.

5.2 Main Results

Table 1 presents the main results on UNISRM-BENCH. We compare four groups of approaches: objective metrics, proprietary AudioLLM models, open-source AudioLLM models, and our proposed UniSRM.

Across all tasks, UniSRM attains the best overall performance. Notably, UniSRM exhibits a larger advantage on context-dependent evaluations (T3/T4), where the judge must integrate textual context or multi-turn conversational context rather than relying solely on local acoustic cues. This observation further suggests that, under rich textual or spoken contextual conditions, many existing models still struggle to deliver reliable judgments. For the MOS-style assessment in T2, UniSRM shows the most consistent alignment with annotated quality scores, providing additional evidence for the effectiveness of our proposed training and optimization strategy.

5.3 Ablations

Table 2 analyzes the effect of our training methods. We include two key variants: *w/o GRPO*, which removes reinforcement learning and keeps only SFT; and *w/o RCR-GRPO*, which still applies GRPO but uses only an accuracy-based reward that measures the correctness of the final decision, without our reasoning-consistent rewards (RCR-GRPO).

Overall, incorporating GRPO consistently improves performance over SFT-only (*w/o GRPO*) training, indicating that on-policy optimization better aligns the judge with evaluation objectives beyond supervised imitation. More importantly, comparing *w/o RCR-GRPO* against our method shows

Model	English				Chinese			
	Text	Scenario	Naturalness	AVG	Text	Scenario	Naturalness	Avg
UniSRM(Ours)	87.45	85.61	81.00	84.69	86.76	91.11	88.14	88.67
w/o RCR-GRPO	89.30	80.81	76.20	82.10	84.78	81.03	78.66	81.49
w/o GRPO	87.64	67.16	63.84	72.88	85.38	70.75	68.97	75.03

Table 5: Multi-dimensional results on the scenario-aware speech preference task. Results are reported as accuracy (↑, %).

Model	Intent	Sim	Context	Emotion	Naturalness	Avg
UniSRM(Ours)	86.51	72.22	83.33	88.89	88.89	83.97
w/o RCR-GRPO	68.25	57.14	67.46	65.87	68.25	65.39
w/o GRPO	69.15	53.17	69.05	50.79	61.90	60.81

Table 6: Multi-dimensional results on the dialogue preference task. Results are reported as accuracy (↑, %).

that adding RCR further improves results across tasks. This suggests that optimizing only the final accuracy is insufficient: without RCR-GRPO, the model may obtain high outcome reward on some samples via shortcut behaviors, while producing rationales that are weakly grounded in the provided context or even inconsistent with the final choice, which in turn degrades overall reliability and performance.

5.4 Fine-grained Analysis Across Tasks and Dimensions

Tables 3–6 report a dimension-level breakdown for all four tasks, offering a closer look at *where* the improvements come from and *how* Reasoning-Consistent Rewards (RCR) shape the reasoning behavior. Overall, UniSRM demonstrates strong and well-balanced performance across nearly all dimensions, rather than improving a single metric at the expense of others.

Task 1. UniSRM improves the preference reliability by consistently strengthening multiple complementary factors, including text fidelity, speaker similarity, expressiveness, and naturalness. Importantly, compared to *w/o RCR-GRPO*, our UniSRM shows more stable gains on the harder perceptual dimensions (e.g., naturalness), which indicates RCR helps maintain a more correct comparison across dimensions.

Task 2. For MOS-style assessment, UniSRM achieves the most consistent alignment with human-annotated aspect scores across the seven criteria. This indicates that RCR is not only beneficial for pairwise judgments but also improves calibration for multi-aspect scoring. In contrast, *w/o RCR-GRPO* tends to underperform across several aspects, suggesting that optimizing only an outcome-based reward provides insufficient guidance for fine-grained scoring behavior and may distort the internal assessment criteria.

Task 3. Table 5 further highlights the advantage of RCR in context-dependent evaluation. UniSRM with RCR yields consistently better performance on scenario-related dimensions in both English and Chinese.

Task 4. For dialogue-conditioned judging, UniSRM also shows robust improvements across different aspects. These dimensions are tightly coupled and require long-context integration over the spoken dialogue history. Our UniSRM provides the largest benefit by encouraging dimension-wise assessments to remain consistent with the final decision.

When accuracy-only GRPO can be worse than SFT. A noteworthy finding across Tables 3–6 is that *w/o RCR-GRPO* (Accuracy-only GRPO) sometimes underperform *w/o GRPO* (SFT-only) on certain dimensions. This phenomenon suggests that optimizing with only an accuracy-based reward may cause the model's reasoning to drift: the policy can exploit shortcuts to obtain favorable outcome rewards while degrading dimension-level judgments, which ultimately hurts overall reliability. Our RCR-GRPO mitigates this issue by directly supervising dimension-wise consistency, leading to more stable improvements across tasks and dimensions.

Model	BVCC		SOMOS-Clean		SOMOS-Full	
	pcc ↑	*acc* ↑	*pcc* ↑	*acc* ↑	*pcc* ↑	*acc* ↑
DNSMOS	0.2990	–	0.0479	–	0.0528	–
Qwen2.5-Omni-7B	0.2563	25.57	0.1561	23.17	0.1484	22.70
Gemini-2.5-Flash	0.3420	29.84	0.2498	29.06	0.2156	27.83
Gemini-2.5-Pro	0.3390	27.42	0.2009	30.71	0.2218	33.94
UniSRM	**0.4977**	**49.16**	**0.2612**	**41.70**	**0.2347**	**52.97**

Table 7: Cross-dataset generalization on speech quality datasets.

5.5 Cross-dataset generalization on speech quality datasets

To validate the generalization ability of UniSRM, we further evaluate it on external human-labeled speech quality datasets, including BVCC (Cooper and Yamagishi, 2021) and SOMOS (Maniati et al., 2022), where SOMOS is entirely unseen during training. Following prior MOS prediction settings, we report both PCC with the human MOS scores and ACC after discretizing MOS into integer bins. For SOMOS, whose ground-truth MOS is fractional, we keep the original decimal scores for PCC and round them only when computing ACC.

As shown in Table 7, these results demonstrate that UniSRM generalizes well beyond the training distribution and is competitive with, or better than, Gemini-family baselines on these human-annotated benchmarks. This finding suggests that UniSRM is not simply overfitting LLM-generated labels, but instead learns reward signals that transfer effectively to unseen domains under human supervision.

6 Conclusion

In this work, we aim to develop a unified speech reward model that can provide multi-dimensional, interpretable judgments for speech evaluation. To this end, we introduce UNISRM-DATA and UNISRM-BENCH, which jointly cover speech evaluation tasks ranging from *utterance-level quality* to *context-level coherency*. Building on this unified data formulation, we present UniSRM with a two-stage pipeline, enabling a single model to support both pairwise preference decisions and fine-grained scoring with explicit reasoning traces. Furthermore, we propose RCR-GRPO to directly supervise dimension-wise reasoning during RL, improving the reliability of rationales. Experimental results on UNISRM-BENCH demonstrate that UniSRM yields more accurate, human-aligned, and robust judgments across all tasks.

Ethics Statement

All models and datasets used in this paper are employed in compliance with their ethical guidelines and licensing terms. When we synthesize speech using open-source TTS systems, we also use these models under their respective licenses. We have provided a clear description of data sources and will release UniSRM-Data and UniSRM-Bench under a suitable open license for research use. Human annotations are conducted by annotators who receive clear instructions, qualification checks, and fair compensation upon completion. We do not collect personally identifiable information during the annotation process. Following common practice, we employ LLMs for semi-automated annotation assistance and model evaluation to improve efficiency and consistency.

Limitations

Despite promising results, our work has several limitations. For example, our current benchmark coverage is limited for challenging scenarios such as heavy accents and overlapped speech. Extending UNISRM-DATA and UNISRM-BENCH to broader acoustic conditions and application settings remains an important direction for future work. Moreover, training and inference with speech-LLM backbones, multi-sample rollouts, and GRPO-based optimization incur non-trivial computational cost. This may limit scalability to larger backbones, larger rollout size G, or broader datasets, and may hinder low-latency deployment when UNISRM is used as an online speech judge. Future work could explore more efficient architectures, lightweight distillation, and caching strategies to reduce inference overhead.

Despite these limitations, we believe UniSRM provides a valuable and practical foundation for the community: an interpretable, multi-dimensional, and unified speech reward modeling framework, together with a comprehensive dataset and benchmark, which can facilitate more reliable reward-based evaluation and optimization for speech generation systems.

Acknowledgement

This work is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 14202623) and National Natural Science Foundation of China (62076144).

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. 2025a. Audio large language models can be descriptive speech quality evaluators. *arXiv preprint arXiv:2501.17202*.

Jingyi Chen, Ju Seung Byun, Micha Elsner, Pichao Wang, and Andrew Perrault. 2025b. Fine-tuning text-to-speech diffusion models using reinforcement learning with human feedback. In *Proc. Interspeech 2025*, pages 3454–3458.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025c. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Erica Cooper and Junichi Yamagishi. 2021. How do voices from past speech synthesis challenges compare today? In *Proc. SSW 2021*, pages 183–188.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Ruibo Fu, Xin Qi, Zhengqi Wen, Jianhua Tao, Tao Wang, Chunyu Qiang, Zhiyong Wang, Yi Lu, Xiaopeng Wang, Shuchen Shi, and 1 others. 2024. Asrrl-tts: Agile speaker representation reinforcement learning for text-to-speech speaker adaptation. *arXiv preprint arXiv:2407.05421*.

Yuan Ge, Junxiang Zhang, Xiaoqian Liu, Bei Li, Xiangnan Ma, Chenglong Wang, Kaiyang Ye, Yangfan Du, Linfeng Zhang, Yuxin Huang, and 1 others. 2025. Sagelm: A multi-aspect and explainable large language model for speech judgement. *arXiv preprint arXiv:2508.20916*.

Yuchen Hu, Chen Chen, Siyin Wang, Eng Siong Chng, and Chao Zhang. 2024. Robust zero-shot text-to-speech synthesis with reverse inference optimization. *arXiv preprint arXiv:2407.02243*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

ITU-T. 1996. Recommendation p.800: Methods for subjective determination of transmission quality (08/96).

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, and 1 others. 2025. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint arXiv:2505.09558*.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. *Preprint*, arXiv:2305.18802.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2022. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. *Preprint*, arXiv:2207.01063.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, and 1 others. 2025. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*.

Potsawee Manakul, Woody Haosheng Gan, Michael J Ryan, Ali Sartaz Khan, Warit Sirichotedumrong, Kunat Pipatanakul, William Held, and Diyi Yang. 2025. Audiojudge: Understanding what works in large audio model based speech evaluation. *arXiv preprint arXiv:2507.12705*.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros

Tsiakoulis. 2022. Somos: The samsung open mos dataset for the evaluation of neural text-to-speech synthesis. *Interspeech 2022*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. 2025. F5r-tts: Improving flow-matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, and 1 others. 2025a. Speechllm-as-judges: Towards general and interpretable speech quality evaluation. *arXiv preprint arXiv:2510.14664*.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. 2025b. Qualispeech: A speech quality assessment dataset with natural language reasoning and descriptions. *arXiv preprint arXiv:2503.20290*.

Xihuai Wang, Ziyi Zhao, Siyu Ren, Shao Zhang, Song Li, Xiaoyu Li, Ziwen Wang, Lin Qiu, Guanglu Wan, Xuezhi Cao, and 1 others. 2025c. Audio turing test: Benchmarking the human-likeness of large language model-based text-to-speech systems in chinese. *arXiv preprint arXiv:2505.11200*.

Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. 2025d. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. 2024a. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*.

Yuanyuan Wang, Hangting Chen, Dongchao Yang, Weiqin Li, Dan Luo, Guangzhi Li, Shan Yang, Zhiyong Wu, Helen Meng, and Xixin Wu. 2025e. Unisep: Universal target audio separation with language models at scale. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Yuanyuan Wang, Hangting Chen, Dongchao Yang, Zhiyong Wu, and Xixin Wu. 2025f. Audiocomposer: Towards fine-grained audio generation with natural language descriptions. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yuanyuan Wang, Hangting Chen, Dongchao Yang, Jianwei Yu, Chao Weng, Zhiyong Wu, and Helen Meng. 2024b. Consistent and relevant: Rethink the query embedding in general sound separation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 961–965. IEEE.

Yuanyuan Wang, Dongchao Yang, Yiwen Shao, Hangting Chen, Jiankun Zhao, Zhiyong Wu, Helen Meng, and Xixin Wu. 2026. Dualspeechlm: Towards unified speech understanding and generation via dual speech token modeling with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 33728–33736.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Dongchao Yang, Haohan Guo, Yuanyuan Wang, Rongjie Huang, Xiang Li, Xu Tan, Xixin Wu, and Helen Meng. 2024a. Uniaudio 1.5: Large language model-driven audio codec is a few-shot audio task learner. *Advances in Neural Information Processing Systems*, 37:56802–56827.

Dongchao Yang, Rongjie Huang, Yuanyuan Wang, Haohan Guo, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. 2025a. Simplespeech 2: Towards simple and efficient text-to-speech with flow-based scalar latent transformer diffusion models. *IEEE Transactions on Audio, Speech and Language Processing*.

Dongchao Yang, Songxiang Liu, Haohan Guo, Jiankun Zhao, Yuanyuan Wang, Helin Wang, Zeqian Ju, Xubo Liu, Xueyuan Chen, Xu Tan, and 1 others. 2025b. Almtokenizer: A low-bitrate and semantic-rich audio codec tokenizer for audio language modeling. In *International Conference on Machine Learning*, pages 70850–70872. PMLR.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Jiang Bian, Zhou Zhao, and 1 others. 2024b. Uniaudio: Towards universal audio generation with large language models. In *Forty-first International Conference on Machine Learning*.

Dongchao Yang, Yuanyuan Wang, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. 2026. Uniaudio 2.0: A unified audio language model with text-aligned factorized audio tokenization. *arXiv preprint arXiv:2602.04683*.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2023. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. 2025c. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*.

Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2024. Speechalign: Aligning speech generation to human preferences. *Advances in Neural Information Processing Systems*, 37:50343–50360.

Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, and 1 others. 2025. Speechjudge: Towards human-level judgment for speech naturalness. *arXiv preprint arXiv:2511.07931*.

Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosongcao Maosongcao, Jiaqi Wang, Weiyun Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, and 1 others. 2025. Omnialign-v: Towards enhanced alignment of mllms with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18490–18515.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Prompt for UniSRM

This section presents the detailed prompts used by UniSRM across all tasks. As illustrated in Figures 3, 4, 5, and 6, we provide the complete system prompt templates for Task 1–4, respectively. Figure 7 shows the prompt used to query GPT-4.1 for constructing the scenario and textual context in Task 3.

Prompt for Speech Evaluation

You are an expert speech evaluator.

Inputs:

[Reference Text]
only for checking text fidelity.

[Prompt Speech]
only for speaker identity similarity.

[Speech A, B]
candidate speech to evaluate.

Your job:

1. Score Speech A and B on FOUR dimensions (0–10 each):
   - (1) Text Fidelity & Intelligibility
   - (2) Speaker Similarity to Prompt Speech
   - (3) Prosody & Expressiveness
   - (4) Naturalness & Audio Quality
2. For Speaker Similarity, use ONLY voice cues (timbre, pitch, accent, style, etc.), not text content.
3. Compute Total_A and Total_B (no ties allowed).
4. Decide which speech is better overall.

Hard constraints:

- In <think>: include scores, explanations, and a [Comparison summary] (2–4 sentences).
- In <answer>: output EXACTLY *"Speech A is better"* or *"Speech B is better"*.

Output format:

```
<think>
[Speech A]
1) Text Fidelity & Intelligibility: score=a1/10; explanation: ...
2) Speaker Similarity to Prompt Speech: score=a2/10; explanation: ...
3) Prosody & Expressiveness Appropriateness: score=a3/10; explanation: ...
4) Naturalness & Audio Quality: score=a4/10; explanation: ...
Total_A = a1+a2+a3+a4 = A_total
[Speech B]
Similar to Speech A.
[Comparison summary]
- 2-4 sentences explaining the main differences and why the winner is better.
</think>
<answer>Speech A is better</answer>
```

Figure 3: Prompt template used for Task 1 (utterance-level speech A/B preference judgment).

Prompt for Speech Quality Assessment (QualiSpeech)

You are an expert judge for speech quality assessment.

Inputs:

[Speech]
one speech sample to evaluate.

Aspects (1–5 each; 1=worst, 5=best):

1. Noise
2. Distortion
3. Speed (speaking rate)
4. Continuity (smoothness / discontinuity)
5. Naturalness
6. Listening effort
7. Overall quality

Your job:

1. Carefully listen to the audio and analyze its quality across all seven aspects.
2. In <think>, first restate concise aspect descriptions (noise / distortion / unnatural pauses / feeling of voice), then provide a coherent paragraph explaining your overall quality judgment in natural language.
3. In <answer>, output ONLY the final scores for all seven aspects in a fixed key=value format.

Hard constraints:

- Scores `N, D, S, C, Na, L, O` MUST be integers in `[1,5]`.
- Use ONLY `<think>...</think>` and `<answer>...</answer>`. No extra text.

Output format (STRICT):

```
<think>
[Aspect descriptions]
Noise description: ...
Distortion description: ...
Unnatural pause: ...
Feeling of voice: ...

[Natural language description]
A detailed paragraph explaining the perceived quality, covering all aspects.
</think>
<answer>noise=N; distortion=D; speed=S; continuity=C; naturalness=Na; listening_effort=L; overall=O;</answer>
```

Figure 4: Prompt template used for Task 2 (speech quality assessment with seven MOS-like aspects).

Prompt for Scenario-Aware Speech Evaluation (EN)

You are an expert judge for SCENARIO-AWARE speech evaluation.

Inputs:

[Scene Context]
Scenario Description, Paragraph Context, Target Emotion.

[Target Text]
the exact sentence that should be spoken.

[Speech A, B]
two audios for the same target text.

Your job:

1. Evaluate Speech A and Speech B as realizations of

the target text under the given context.
2. Score each speech on THREE dimensions (0–10 each) with 1–2 sentence explanations:
   - (1) Text Fidelity & Intelligibility
   - (2) Scenario Style Match **[CRITICAL]**
   - (3) Naturalness & Audio Quality
3. Compute Total_A and Total_B as the sum of the three scores (they MUST be different).
4. In <answer>, decide which speech is better overall.

Dimension hints:
- **Text Fidelity & Intelligibility:** matches the target text; clear and understandable.
- **Scenario Style Match:** emotion and speaking style fit the target emotion and context.
- **Naturalness & Audio Quality:** human-like, stable, and comfortable to listen to.

Hard constraints:
- Output ONLY <think> and <answer>, nothing else.
- In <think>: include both [Speech A] and [Speech B] with scores and explanations, and a [Comparison summary] (2–4 sentences).
- In <answer>: output EXACTLY *"Speech A is better"* or *"Speech B is better"*.

Output format:

```
<think>
[Speech A]
1) Text Fidelity & Intelligibility: score=a1/10; explanation: ...
2) Scenario Style Match: score=a2/10; explanation: ...
3) Naturalness & Audio Quality: score=a3/10; explanation: ...
Total_A = a1+a2+a3 = A_total
[Speech B]
Similar to Speech A.
[Comparison summary]
- 2-4 sentences highlighting the main differences and why the winner is better.
</think>
<answer>Speech A is better</answer>
```

Figure 5: Prompt template used for Task 3 (Scenario-aware evaluation, EN).

Prompt for Multi-turn Spoken Dialogue Evaluation

You are an expert judge for multi-turn SPOKEN dialogues.

Inputs:

[dialog_history]
audios containing all previous turns, up to but NOT including the current turn.

[speech_A]
candidate response A for the current turn.

[speech_B]
candidate response B for the current turn.

Your job:
1. Evaluate both candidates A and B as possible next turns given `dialog_history`.
2. Score each candidate on FIVE dimensions (0–10 each) with 1–2 sentence explanations:
   - (1) Intent Matching & Dialogue Act
   - (2) Speaker Consistency
   - (3) Contextual Consistency
   - (4) Emotion & Prosody Match
   - (5) Overall Naturalness
3. Compute the total score for Speech A and Speech B (sum of the five dimensions; totals MUST be different), then decide which speech is better overall.

Dimension hints:
- **Intent Matching & Dialogue Act:** does the reply follow the topic and intent appropriately?
- **Speaker Consistency:** does the voice match the same person in relevant turns (timbre, pitch, gender cues, accent, speaking style)?
- **Contextual Consistency:** is the content consistent with previous turns without contradictions?
- **Emotion & Prosody Match:** is the emotion, tone, and prosody suitable for the current situation?
- **Overall Naturalness:** does it sound like a coherent, natural human reply in this dialogue?

Hard constraints:
- Output ONLY <think> and <answer>, nothing else.
- In <answer>: output EXACTLY *"Speech A is better"* or *"Speech B is better"*.

Output format (STRICT):

```
<think>
[Speech A evaluation]
- Intent Matching & Dialogue Act: score=a1/10; explanation: ...
- Speaker Consistency: score=a2/10; explanation: ...
- Contextual Consistency: score=a3/10; explanation: ...
- Emotion & Prosody Match: score=a4/10; explanation: ...
- Overall Naturalness: score=a5/10; explanation: ...
- Total score for Speech A = a1+a2+a3+a4+a5 = A_total
[Speech B evaluation]
Similar to Speech A.
[Comparison summary]
- Brief comparison of Speech A vs Speech B across the five dimensions (2-4 sentences), explaining why the chosen speech fits the multi-turn context better.
</think>
<answer>Speech A is better</answer>
```

Figure 6: Prompt template used for Task 4 (multi-turn dialogue evaluation).

Prompt for Scenario Context Construction using GPT-4.1.

You are a skilled scene and story generator for speech data.

Inputs:

[Utterance Text]
a single sentence that will be spoken.

[Emotion Label]
the target emotion of how this sentence is spoken (e.g., Neutral, Angry, Happy, Sad, Surprise).

[Target Language]
the language of the utterance.

Your job:

1. Construct a coherent scenario and short story context where the utterance would naturally appear.
2. Ensure the scenario and context make the given emotion label reasonable and consistent.
3. Ensure the context logically leads to the utterance text.

Hard constraints:

- Output MUST be a strict JSON object with exactly two fields: `scenario_description` and `paragraph_context`.
- The language of both fields MUST be {LANG}.
- Do NOT rewrite or change the utterance text itself.
- Make the emotion expression implicitly reasonable; avoid explicitly stating the emotion in every sentence (e.g., do not repeatedly say "he is angry"), but ensure the situation reflects the target emotion.

Output format (STRICT JSON):

```
{
"scenario_description": "...",
"paragraph_context": "..."
}

[Utterance Text]: "{UTT_TEXT}"
[Emotion Label]: "{EMOTION}"
[Target Language]: "{LANG}"
```

Figure 7: Prompt of generating scenario context conditioned on text content and emotion label for scenario-aware evaluation.

B Task-specific Evaluation Dimensions

Table 8 summarizes the evaluation dimensions required by each task in our dataset. Preference-based tasks output per-dimension scores along with a final A/B decision, while the MOS-style task outputs seven aspect scores.

C Case Study: Example Output

Figures 8–11 present example responses of UniSRM on Task 1 to Task 4, respectively. The example illustrates how the model conducts a comprehensive multi-dimensional evaluation across different key aspects. As shown in the detailed reasoning process, the model provides interpretable, fine-grained assessments for each dimension before aggregating the scores to reach a final comparative judgment.

D Model Configuration

We use a learning rate of 1.0×10^{-5} with gradient accumulation of 8 steps for SFT, and a learning rate of 1.0×10^{-6} with gradient accumulation of 2 steps for RL. During GRPO, we sample $G = 8$ completions per prompt. We set the reward weights to $\lambda_{\text{fmt}} = \lambda_{\text{acc}} = \lambda_{\text{rc}} = 1$ when computing the total reward, and use a KL coefficient of $\beta = 0.04$. For reproducibility, we summarize the main training hyperparameters of both stages in Table 10. We adopt representative open-source TTS systems as speech synthesizers, including CosyVoice2 (Du et al., 2024), F5-TTS (Chen et al., 2025c), ChatTTS[2], and XTTS[3].

Objective metrics such as WER, SIM, UTMOS, and DNSMOS are substantially faster to compute, but each captures only a narrow aspect of speech quality and cannot serve as a unified, context-aware evaluator. For transparency, we also report the computational cost of UniSRM. At inference time, UniSRM runs at 8.98 seconds per iteration with approximately 20 GB peak GPU memory, which is comparable to SpeechJudge under the same hardware and input settings. For training, the SFT stage takes about 4 hours for one epoch on 8 GPUs, corresponding to 30.94 GPU-hours, with roughly 40 GB peak memory per GPU. The GRPO stage takes about 60 hours, corresponding to 480 GPU-hours, with roughly 30 GB peak memory per GPU.

E Dataset statistics

Building on the data introduction in Section 3, Figure 13 presents the detailed pipeline of UniSRM dataset construction across all tasks. Task 1 uses a large-scale speech preference dataset constructed from the LibriTTS-R (Koizumi et al., 2023) corpus. Task 2 is built upon the public QualiSpeech (Wang et al., 2025b) dataset for fine-grained MOS-style speech quality assessment. Task 3 is constructed from the Emotional Speech Dataset (ESD) (Zhou et al., 2022) for scenario-aware style consistency under textual context (English and Chinese). Task 4 is based on DailyTalk (Lee et al., 2022) for multi-turn dialogue speech evaluation conditioned on dialogue history. For all tasks, we partition the collected data into SFT, RL, and Bench subsets, ensuring no overlap across splits to avoid leakage. Finally, the size of datasets used in each training stage is as shown in Table 9.

F Evidence Groundedness (EG) Analysis

We additionally report EG_mean to assess the evidence groundedness and reliability of the model's reasoning, since accuracy-only GRPO (w/o RCR-GRPO) may yield decisions with generic or weak justifications. For each sample, we use GPT-4.1

[2] https://github.com/2noise/ChatTTS
[3] https://github.com/coqui-ai/TTS

Task	Dimension	Range
Task 1:Utterance-Level Speech A/B Preference Judgment	Text Fidelity & Intelligibility	0–10
	Speaker Similarity to Prompt Speech	0–10
	Prosody & Expressiveness Appropriateness	0–10
	Naturalness & Audio Quality	0–10
Task 2: Utterance-Level Speech Quality Assessment	Noise	1–5
	Distortion	1–5
	Speed (speaking rate)	1–5
	Continuity (smoothness / discontinuity)	1–5
	Naturalness	1–5
	Listening effort	1–5
	Overall quality	1–5
Task 3: Scenario-Aware Style Coherency conditioned on Textual Context	Text Fidelity & Intelligibility	0–10
	Scenario Style Match	0–10
	Naturalness & Audio Quality	0–10
Task 4: Multi-Turn Dialogue Speech Evaluation conditioned on Dialogue History	Intent Matching & Dialogue Act	0–10
	Speaker Consistency	0–10
	Contextual Consistency	0–10
	Emotion & Prosody Match	0–10
	Overall Naturalness	0–10

Table 8: Task-specific evaluation dimensions used in UNISRM.

to judge reasoning results of w/o RCR-GRPO and our UniSRM, and then assign an EG score in $\{0, 1, 2\}$, where 0 indicates no concrete evidence, 1 indicates some task-relevant evidence but limited linkage, and 2 indicates multiple concrete, observable, task-relevant observations clearly supporting the final decision. We then average the scores over all samples to obtain EG_mean. Table 11 summarizes EG_mean across tasks, where higher values indicate more evidence-grounded rationales with clearer support for the final choice. Overall, UniSRM consistently achieves higher EG_mean, suggesting improved alignment between the reasoning process and the final decision, which can be attributed to the effectiveness of our RCR-GRPO strategy.

G Data Quality Control and Human Verification Criteria

G.1 Data Quality Control

To prevent the reward model from memorizing positional cues, we randomly shuffle the order of (s_A, s_B) for every pair, so that the first and second positions carry no deterministic meaning. Since automatic judgments may still contain noise, we further remove self-contradictory samples by filtering cyclic conflicts, e.g., $A > B$, $B > C$, $C > A$ within the same text group.

To avoid positional bias, we randomly permute the order of (s_A, s_B) for each comparison pair, so that the first and second positions carry no deterministic meaning. Moreover, since LLM-based annotations can be noisy, we perform a consistency-based filtering step within each text group: we discard samples that participate in cyclic contradictions (e.g., $A > B$, $B > C$, yet $C > A$), which typically indicate unreliable or unstable judgments. These procedures, together with the human verification criteria described in Appendix G.2, help improve the overall label reliability for optimization and evaluation.

G.2 Human Verification Criteria

To maximize label reliability for optimization and evaluation, we additionally conduct human verification on the GRPO and Test subsets, retaining only samples aligned with human preference. Before starting the task, each annotator was shown an in-

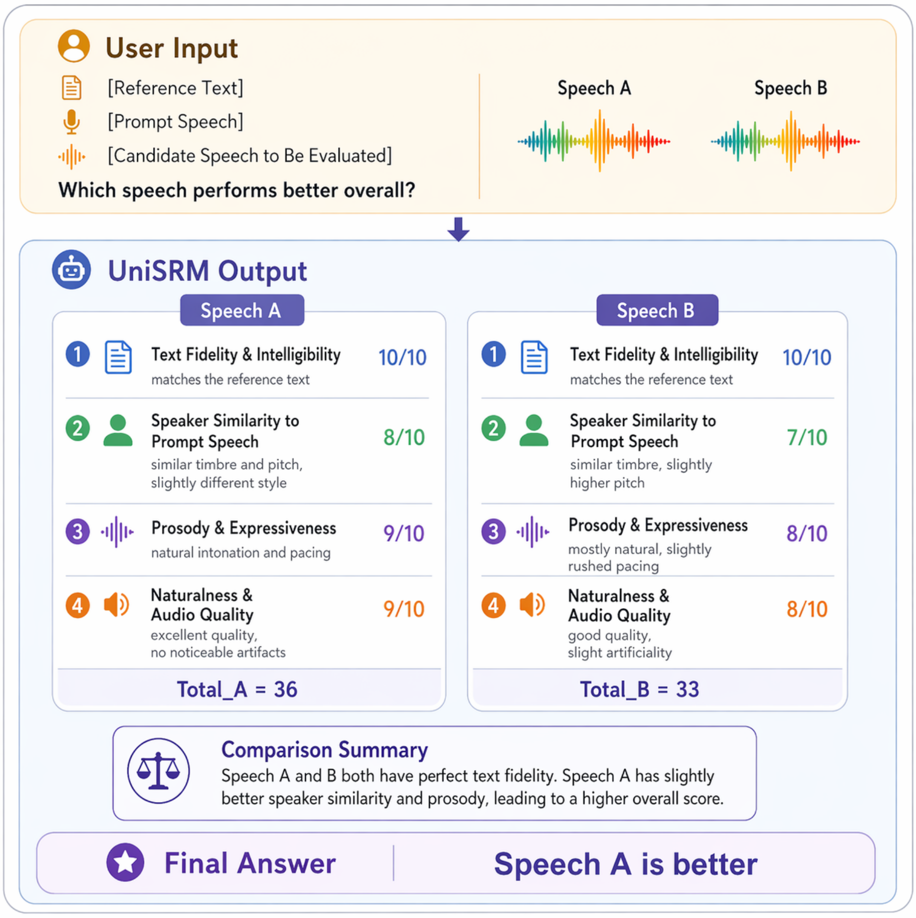

Figure 8: Example output of Task 1 (utterance-level speech A/B preference judgment) in UniSRM.

struction page describing (i) what data they would listen to, (ii) the purpose of the study (training and evaluating speech reward models), and (iii) how the collected labels would be used (research-only optimization and benchmarking). Annotators provided informed consent by explicitly agreeing to the task terms prior to annotation. We paid approximately RMB 1.05 for each pairwise comparison item, i.e., one A/B judgment. Regarding the underlying speech data, we curate speech samples from publicly available or properly licensed datasets; we follow the corresponding licenses/terms of use, and do not collect additional personal information beyond what is already contained in the source datasets.

Specifically, we perform human verification on the GRPO and Test subsets of utterance-level speech A/B preference judgment task to ensure high label reliability for optimization and evaluation. Each item presents the same input condition, including target text, speech prompt (reference speaker) and a candidate pair (s_A, s_B), together with an auto-generated preference label.

Each pair is evaluated by three independent annotators. Annotators must satisfy the following requirements:

- Language proficiency: fluent listening ability in English.
- Equipment and environment: headphones are required; annotation must be conducted in a quiet environment.
- Hearing and attention: no known hearing impairments.
- Qualification: annotators must pass a short qualification test containing gold-reference items before annotation.

For each pair, the interface displays the necessary inputs, including target text, prompt speech, and provides candidate speeches s_A and s_B. Annotators follow the rules below:

- Playback: listen to both samples in full. Replaying is allowed; however, each sample should not be replayed excessively (recommended ≤ 3 times) to avoid fatigue-induced inconsistency.
- Order control: the presentation order of s_A and s_B is randomized per annotator.

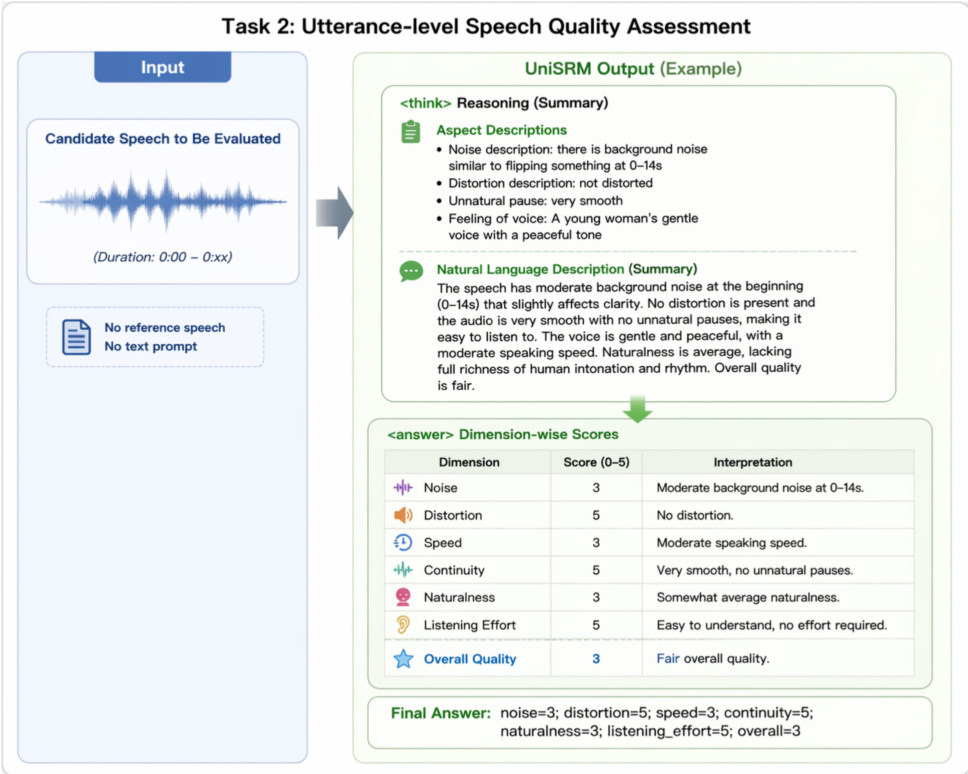

Figure 9: Example output of Task 2 (utterance-level speech quality assessment) in UniSRM.

Task	**SFT**	**RL**	**Bench**	**Total**
Task 1 (A/B Pref.)	11146	2787	498	14431
Task 2 (QualiSpeech MOS)	10558	2167	1852	14577
Task 3 (Scenario, EN)	5018	1815	542	7375
Task 3 (Scenario, ZH)	4869	1989	506	7364
Task 4 (Dialogue)	1470	916	126	2512
Total	**33061**	**9674**	**3524**	**46259**

Table 9: Data statistics (#number) of all UniSRM datasets across tasks. Columns report the sizes of the SFT subset, RL subset, and UNISRM-BENCH.

	Stage 1	Stage 2
Batch size / GPU	1	1
#GPUs	8	8
Gradient accumulation	8	2
Effective batch size	64	16
Learning rate	1×10^{-5}	1×10^{-6}
Precision	bf16	bf16

Table 10: Training configuration of UniSRM.

- Decision: provide a binary preference (*A is better* vs. *B is better*); ties are not allowed.
- Criteria: text fidelity/intelligibility, speaker similarity, prosody/expressiveness, and naturalness/audio quality.
- Invalid cases: if the pair is not comparable due to corruption, severe truncation, missing content, or other fatal issues, mark it as *invalid*.

Model	**Task1**	**Task3-En**	**Task4**
w/o RCR-GRPO	1.28	1.66	1.95
UniSRM (Ours)	1.57	1.90	1.98

Table 11: EG_mean (↑) across tasks.

Majority vote and acceptance criteria. Let $y^{\text{human}} \in \{A, B, \text{invalid}\}$ denote each annotator's decision. We retain a pair if and only if it satisfies all of the following:

Figure 10: Example output of Task 3 (scenario-aware style consistency evaluation) in UniSRM.

- Validity: at least two annotators do *not* mark the pair as invalid.
- Majority agreement: at least 2 out of 3 annotators agree on the preferred sample.
- Consistency with the auto label: the auto-generated preference label by Gemini matches the human majority decision.

Pairs failing any criterion are discarded from the GRPO/Test subsets.

Tie and low-consensus handling. If the three annotators do not yield a strict majority (e.g., one votes A and one votes B and one marks invalid, or A/B split without majority), the pair is removed. This conservative filtering avoids introducing ambiguous supervision during GRPO and ensures a clean evaluation set.

Quality control. Annotators who consistently fail the gold items or exhibit abnormal behavior (e.g., extremely short annotation time, always selecting the left item) are excluded, and their annotations are discarded and replaced with re-collected annotations from qualified annotators.

H Multi-turn dialogue setting

Figure 12 shows an example input for the multi-turn dialogue evaluation task (Task 4). Both UniSRM-Data and UniSRM-Bench contain multi-turn spoken dialogue samples designed to assess context-dependent judgment ability. Each sample includes the full preceding dialogue context and the current response turn to be evaluated. For this task, the dialogue history is provided entirely as raw audio, without transcripts, and the full conversation context is used in both training and evaluation.

Example of a Multi-turn Dialogue Evaluation Instance

System: task definition and evaluation rubric.

User content (audio sequence):

[Turn 1 audio: Speaker A]
"Hi, could you help me book a hotel for tomorrow?"

[Turn 2 audio: Speaker B]
"Sure—what city and budget?"

[Turn 3 audio: Speaker A]
"Singapore, around 150 SGD."

Target to judge:

Figure 11: Example output of Task 4 (multi-turn dialogue evaluation) in UniSRM.

[Turn 4 audio: Speaker B]
the response audio to be evaluated.

Figure 12: Input example of a multi-turn dialogue evaluation instance.

I Baseline input parity

To ensure fair comparisons, we enforce input parity across UniSRM and all judge-style baselines whenever the baseline supports the corresponding modality. The detailed input configuration for each task is summarized in Table 12. For single-turn tasks, all judge models receive the same audio for evaluation and the same textual task instruction. For multi-turn dialogue tasks, all judge models also receive the same raw audio dialogue history, including the full conversation context. We do not provide any other side information, such as transcripts or attribute tags, to UniSRM or any baseline. For objective metrics that inherently do not support textual prompts or dialogue history (e.g., WER, SIM, UTMOS, DNSMOS), we report results under their native audio-only setting.

Method	Audio to Evaluate	User Prompt	Dialogue History
Single-turn tasks (T1/T2/T3)			
Objective Metrics	raw audio	✗	N/A
Proprietary Models	raw audio	text	N/A
Open-source Models	raw audio	text	N/A
UniSRM	raw audio	text	N/A
Multi-turn dialogue task (T4)			
Objective Metrics	raw audio	✗	✗
Proprietary Models	raw audio	text	raw-audio history (full conversation)
Open-source Models	raw audio	text	raw-audio history (full conversation)
UniSRM	raw audio	text	raw-audio history (full conversation)

Table 12: Input parity across UniSRM and baselines. N/A indicates that the corresponding field (e.g., Dialogue History) is not part of the task input. ✗ indicates that objective metrics only support audio-only input and do not take textual prompts.

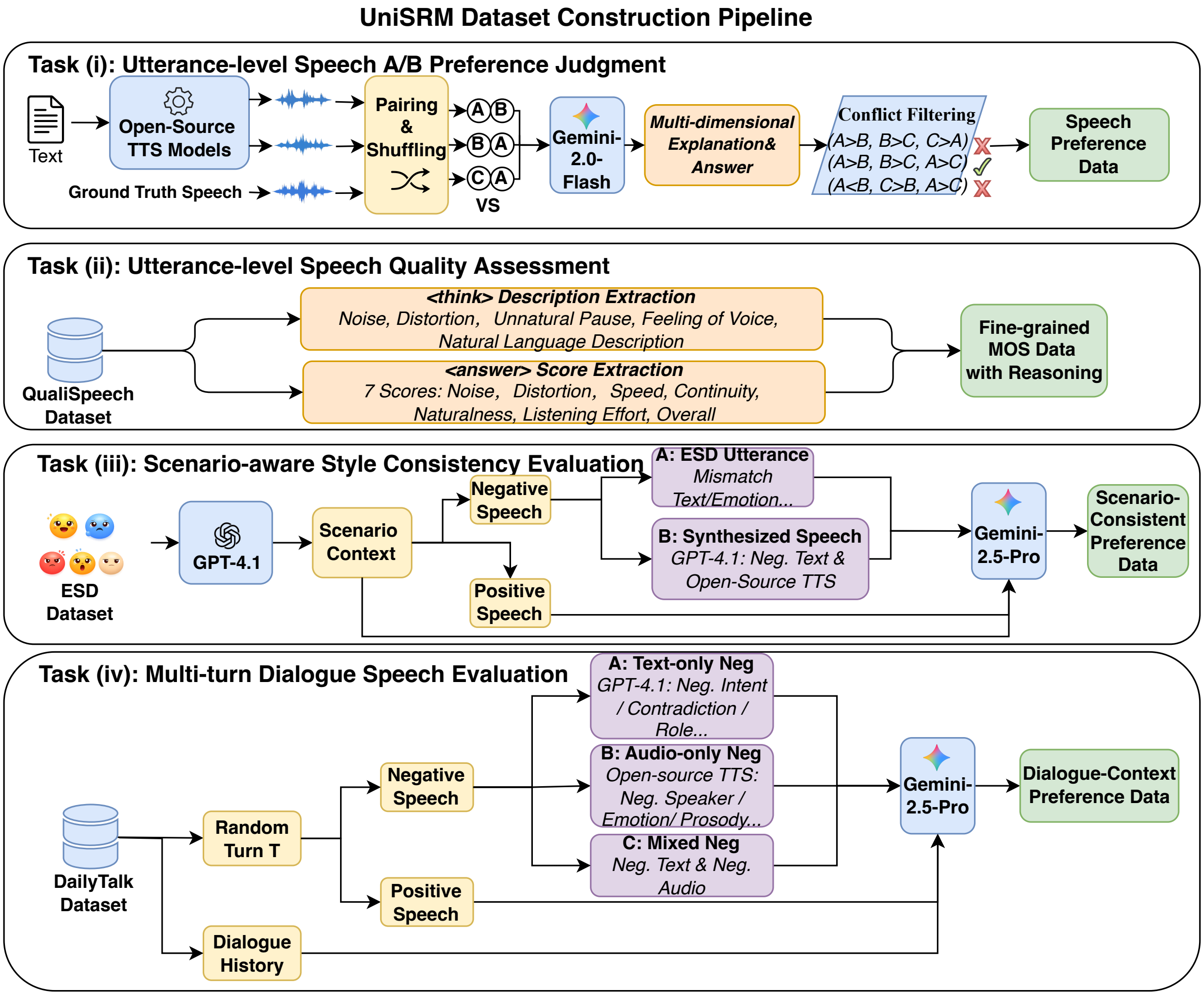

Figure 13: The Detailed Pipeline of UniSRM-Data Construction.