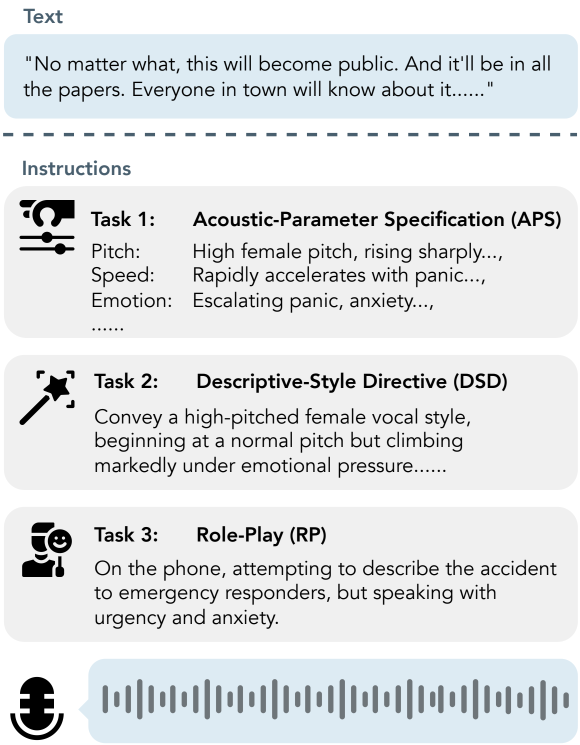
InstructTTSEval: Benchmarking Complex Natural-Language Instruction Following in Text-to-Speech Systems

Anonymous ACL submission

Abstract

In modern speech synthesis, paralinguistic information—such as a speaker's vocal timbre, emotional state, and dynamic prosody—plays a critical role in conveying nuance beyond mere semantics. Traditional Text-to-Speech (TTS) systems rely on fixed style labels or inserting a speech prompt to control these cues, which severely limits flexibility. Recent attempts seek to employ natural-language instructions to modulate paralinguistic features, substantially improving the generalization of instruction-driven TTS models. Although many open-source and commercial systems now support customized synthesis via textual description, their actual ability to interpret and execute complex instructions remains largely unexplored. In addition, there is still a shortage of high-quality benchmarks and automated evaluation metrics specifically designed for instruction-based TTS, which hinders accurate assessment and iterative optimization of these models. To address these limitations, we introduce **InstructTTSEval**, the *first* TTS benchmark for measuring the capability of complex natural-language style control. InstructTTSEval includes three tasks, namely Acoustic-Parameter Specification, Descriptive-Style Directive, and Role-Play, including English and Chinese subsets, each with 1k test cases (6k total) paired with reference audio. We leverage Google's Gemini as an automatic judge to assess their instruction-following abilities. Our evaluation of accessible instruction-following TTS systems reveals that even the best-performing model achieves only modest style-control accuracy, underscoring substantial room for improvement. We anticipate that InstructTTSEval will drive progress toward more powerful, flexible, and accurate instruction-following TTS models. [1]

1 Introduction


Text
Instructions
Task 1: Acoustic-Parameter Specification (APS)
Pitch:
Speed:
Emotion:
Task 2: Descriptive-Style Directive (DSD)
Task 3: Role-Play (RP)

"No matter what, this will become public. And it'll be in all the papers. Everyone in town will know about it......"

High female pitch, rising sharply...,
Rapidly accelerates with panic...,
Escalating panic, anxiety...,
......

Convey a high-pitched female vocal style, beginning at a normal pitch but climbing markedly under emotional pressure......

On the phone, attempting to describe the accident to emergency responders, but speaking with urgency and anxiety.

Figure 1: Tasks in InstructTTSEval, progressing from concrete control to abstract expressiveness. APS task evaluates models' accurate control for all low-level acoustic features, DSD task tests a model's ability to generalize from unstructured prompts, and RP task requires models to infer appropriate vocal styles from high-level character or scenario descriptions.

In recent years, a number of standout TTS systems have emerged (Ye et al., 2025; Wang et al., 2024b; Liao et al., 2024; Ren et al., 2020; Chen et al., 2024; Betker, 2023; Wang et al., 2023, 2025; Anastassiou et al., 2024), achieving extremely low word- or character-error rates, highly natural "human-like" fluency, and remarkable voice-cloning abilities.

Despite their strong semantic clarity, human conversation carries vital acoustic cues as well (Cutler et al., 1997). In an extreme example, the same text spoken by different people—or by the same person in different moods—can convey entirely different meanings. For instance, when someone says

[1] We will release our data soon after acceptance.

"Help me", a playful child's request might simply mean "help me reach that toy" or "tie my shoe," delivered with innocent enthusiasm rather than real distress; an elderly speaker, however, could express genuine urgency—"I need assistance because I'm injured or unsteady"—signaling true vulnerability rather than casual plea. Controlling such acoustic features is crucial even for modern TTS systems: we would not expect comforting words rendered in a cold, detached tone by a speech engine, as that mismatch could largely diminish the quality of the user's experience.

There have been initial attempts to control acoustic features—most use special tokens (e.g., <happy>) or short phrase prompts (Du et al., 2024a; Guo et al., 2024; Kim et al., 2021), and some recent works explore free-form natural-language control and show encouraging progress (Yang et al., 2023; Guo et al., 2022; Leng et al., 2023; Liu et al., 2023; Du et al., 2024b; Ji et al., 2023, 2024; Lyth and King, 2024; Zhou et al., 2024; Yang et al., 2025). However, current metrics are insufficient for evaluation: most common objective metrics measure speech quality like word error rate (WER) and speaker similarity (SIM), while subjective MOS evaluations depend on costly human annotation and often suffer from inconsistent standards. We still lack standardized benchmarks specifically designed to quantify the effectiveness of natural-language instruction-based acoustic control, hindering accurate assessment and iterative model improvement.

To address this gap, we introduce **INSTRUCTTTSEVAL**, a fully automatic benchmark for measuring a TTS system's ability to control acoustic features. As shown in Fig. 1, it consists of three tasks: 1) Acoustic-parameter Specification: models receive a structured set of fine-grained natural-language instructions specifying detailed acoustic characteristics (e.g., pitch, speed, emotion), and must directly map each descriptive cue to the corresponding acoustic realization. 2) Descriptive-style Directive: models receive more open-ended, qualitative style instructions expressed freely in natural language. It must parse this holistic description and infer the underlying parameter adjustments (e.g., prosody, speed, intensity) needed to produce the requested expressive style. 3) Role-play: models are given abstract, high-level role and scenario descriptions—closer to the kinds of prompts non-expert users might provide—and must leverage their own knowledge and contextual understanding to infer coherent acoustic expressions (emotion, volume, tone, etc.) and manifest these choices in the synthesized output.

To ensure realism, we build our dataset bottom-up: we mine highly expressive clips from movies and TV, apply rigorous cleaning and filtering, and then reverse-generate style instructions from the audio. In total, we offer three tasks with 1,000 English and 1,000 Chinese examples each (6,000 instructions overall), each paired with a reference wave collected data. Finally, we leverage Gemini's powerful speech-understanding capabilities—using an LLM-as-a-judge setup—to evaluate today's state-of-the-art instruct TTS systems. Results show that even the top-scoring model only achieved a 71.1 in the EN subset and 51.1 in the ZH subset, highlighting that fine-grained acoustic control remains a major open challenge. Meanwhile, our case studies reveal the significant shortcomings in current TTS systems when it comes to reproducing natural vocal events, handling extreme emotional transitions, and synthesizing character-specific timbres—capabilities that are crucial for advancing TTS toward truly human-like expressiveness.

In summary, our contributions are as follows:

- We propose INSTRUCTTTSEVAL, the *first* automatic benchmark for instruction-following TTS, comprising hierarchical tasks to comprehensively evaluate a model's ability to interpret and execute complex natural-language style descriptions.
- Confirming strong agreement with human annotations, we leverage Gemini as a judge to conduct rapid, scalable automatic assessment.
- We benchmark a diverse collection of popular open-source and commercial TTS systems on INSTRUCTTTSEVAL, providing detailed analysis for future model improvements.

2 Related Work

2.1 Controllable TTS

In previous work, researchers have taken several different approaches to controlling acoustic features in TTS. CosyVoice (Du et al., 2024a) and FireRedTTS (Guo et al., 2024) seek to insert special tokens into the input to steer the generated speech. ST-TTS (Kim et al., 2021) similarly uses dedicated style tags to modulate prosody and timbre.

Meanwhile, models such as InstructTTS (Yang et al., 2023), PromptTTS series (Guo et al., 2022;

Dataset/Benchmark	# Labels	Highlights	Annotation	Hier.
TextrolSpeech (Ji et al., 2023)	5	-	Fixed	✗
PromptSpeech (Guo et al., 2022)	5	-	Fixed	✗
SpeechCraft (Jin et al., 2024)	8	Emphasis	Fixed	✗
ParaSpeechCraft (Diwan et al., 2025)	11	Sound event	Fixed	✗
INSTRUCTTTSEVAL (ours)	12	Emphasis, sound event, dynamic changes	Free-form	✓

Table 1: Comparison with existing open-sourced style description datasets. "Hier." denotes hierarchical design. "Fixed" annotation refers to labels drawn from a limited set of tags or classifier outputs; "free-form" indicates natural-language descriptions that vary on a case-by-case basis.. In all cases, the initial annotations are rewritten into fluent style instructions using a language model.

Leng et al., 2023), PromptStyle (Liu et al., 2023), and ParlerTTS(Lyth and King, 2024) allow users to describe desired voice characteristics in free-form text rather than relying on rigid tokens. CosyVoice 2 (Du et al., 2024b) also goes further by permitting a natural-language style prompt before the text. Salle (Ji et al., 2023) combines speech tags with LLM-based style rewriting; and ControlSpeech (Ji et al., 2024) additionally integrates an example speech prompt for guidance. VoxInstruct (Zhou et al., 2024) merges style directives and transcript into a single instruction. And EmoVoice (Yang et al., 2025) focuses specifically on conveying emotional nuance via natural-language descriptions. For commercial offerings, Services like Hume AI [2] and ElevenLabs [3] let users simply type in style descriptions to shape output, and the recent GPT-4o-mini TTS [4] likewise supports free-form prompts for style control. Despite this diversity of techniques, no unified evaluation framework exists to measure and compare their real-world style-control capabilities. To fill this gap, we aim to propose a benchmark designed explicitly to assess how effectively these models follow user-defined style instructions.

2.2 Acoustic-featured Datasets and Benchmarks

Traditional TTS Evaluation Metrics Early TTS research has predominantly measured performance in terms of intelligibility and speaker similarity—most commonly using word error rate (WER) and speaker-similarity (SIM) scores (Panayotov et al., 2015; Anastassiou et al., 2024). While these metrics effectively capture whether the synthesized speech is accurate and natural-sounding, they do not assess a system's ability to follow detailed style or prosody instructions. Meanwhile, subjective metrics such as Mean-Opinion-Score (MOS) rely on human annotation, which is time-consuming and extremely costly.

Speech Understanding Benchmarks Benchmarks focusing on acoustic features mostly include speech understanding tasks. AudioBench (Wang et al., 2024a) and SD-Eval (Ao et al., 2024) integrate voice understanding tasks to assess models' ability to perceive paralinguistic information like accent, gender, and emotion. And Salmon (Maimon et al., 2024) measures whether the model can identify the inconsistencies in the input speech. These suites excel at evaluating recognition and classification, but they do not measure how well a TTS model can produce speech that matches a user-defined acoustic description.

Datasets with Style Description Several recent datasets pair speech samples with natural-language style descriptions using a variety of pipelines. TextrolSpeech (Ji et al., 2023) begins with five discrete features (gender, pitch, speed, volume, emotion) and relies on an LLM to weave those elements into coherent prompts. AudioBox (Vyas et al., 2023) and PromptSpeech (Guo et al., 2022) both have human annotators label clips along core dimensions and then apply an LLM to rewrite those tags into fluent sentences. And NLSpeech (Yang et al., 2023) relies entirely on manually annotated data. generates descriptive keywords (gender, speed, pitch, volume) before forming sentences and retrieving Spoken Language Understanding (SLU)-tagged audio. SpeechCraft (Jin et al., 2024) adopts a bottom-up approach: a classifier first assigns tags to audio (e.g., "elderly", "emphasis on..."), and an LLM then composes those tags into full descriptions. ParaSpeechCraft (Diwan et al., 2025) expands the tag set to 58 labels spanning inherent speaker traits and situational context.

[2] https://www.hume.ai/

[3] https://elevenlabs.io/text-to-speech

[4] https://www.openai.fm/

While each method offers useful insights, they face several challenges: 1) they cannot handle nuanced features in a speech, such as emotion transition; 2) they rely on heavy human annotation, or classifiers that are only available for limited features. We illustrate the major difference between our dataset and previous work in Fig. 1. To address these gaps, we introduce an automatic and scalable pipeline for constructing style captions and instructions, capturing fine-grained acoustic attributes, to enable continuous evaluation and enhancement of controllable TTS systems.

3 INSTRUCTTTSEVAL

3.1 Task Definition

Based on previous studies (Cutler et al., 1997; Diwan et al., 2025; Jin et al., 2024), we integrate 12 features across four tiers: physiological (e.g., gender, pitch, texture), linguistic (e.g., clarity, fluency, speed), social (e.g., accent, age, volume), and psychological or pragmatic (e.g., emotion, tone, personality), as listed in Fig. 3. Building upon these features, we design the following tasks taskscorresponding to three instruction granularities, as illustrated in Fig. 1, to evaluate controllable TTS systems:

1. *Acoustic-Parameter Specification* (APS) focuses on fine-grained control over low-level acoustic attributes. The input consists of explicit instructions covering all 12 features, and the goal is to assess whether the model can independently manipulate each property with precision.

2. *Descriptive-Style Directive* (DSD) is a more naturalistic variant where the structured instructions from APS are rewritten by an LLM into free-form descriptions. We further introduce diversity by randomly omitting some attributes in the prompt. This task examines the model's ability to generalize from unstructured input and produce appropriate speech styles even when certain attributes are unspecified.

3. *Role-Play* (RP) challenges the model's contextual and social reasoning abilities. Instead of explicitly stating vocal traits, the prompts describe roles or scenarios (e.g., a teacher scolding a student, a nervous applicant in an interview). The model is expected to infer the corresponding vocal style based on world knowledge and map abstract social cues to concrete acoustic realizations, such as changes in speed, volume, intonation, or phrasing.

Together, these three tasks offer a comprehensive benchmark for evaluating both low-level controllability and high-level style generalization in controllable TTS systems.

3.2 Data Collection

The overall process of how we construct INSTRUCTTTSEVAL is illustrated in Fig. 2.

3.2.1 Data Preprocessing

Data Source To design prompts that exhibit strong stylistic expression with coherent feature alignment, we curate our data from movies, TV dramas, and variety shows—domains rich in expressive speech and diverse emotional delivery. We utilized the publicly available NCSSD (Liu et al., 2024) dataset and additionally collect and process supplementary audio from various film and television sources.

Data Cleaning For our self-collected data, we apply speaker diarization (Bredin, 2023; Plaquet and Bredin, 2023) to acquire segments shorter than 30 seconds of the same speakers. We then use `whisper-large-v3` (Radford et al., 2022) for automatic speech recognition (ASR), followed by punctuation restoration using `ct-punc` [5] and BELLE (BELLEGroup, 2023). This pipeline yields approximately 6,000 hours of transcribed speech.

Filtering To ensure high audio quality, we filter both NCSSD and our own data using DNSMOS (Reddy et al., 2021) with a threshold of 2.8. We further refine the dataset with a custom-tuned WhisperD (Darefsky et al., 2024) model to reserve single-speaker segments. To improve the accuracy of downstream speech caption, we retain only segments longer than 3 seconds and containing more than 10 words (for English) or 10 characters (for Chinese). To select highly expressive speech samples, we employ the DVA toolkit (Wagner et al., 2022). We retain only the Dominance (Potency–Submissiveness) and Arousal (Activation–Deactivation) dimensions, filtering samples with both scores exceeding a threshold of 0.8. The Valence (V) dimension is discarded, as our benchmark aims to cover both positive and negative emo-

[5] https://huggingface.co/funasr/ct-punc

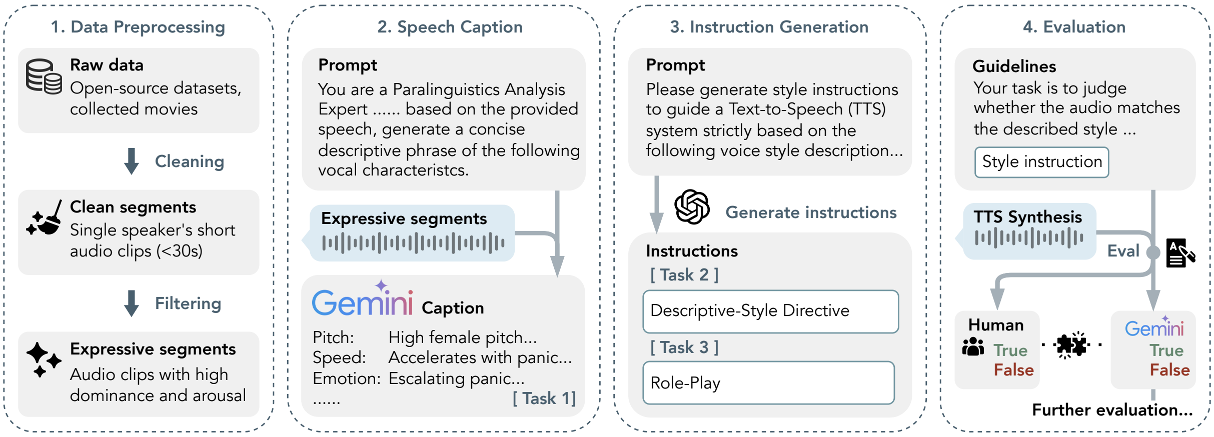

1. Data Preprocessing
Raw data
Open-source datasets, collected movies
Cleaning
Clean segments
Single speaker's short audio clips (<30s)
Filtering
Expressive segments
Audio clips with high dominance and arousal
2. Speech Caption
Prompt
You are a Paralinguistics Analysis Expert ...... based on the provided speech, generate a concise descriptive phrase of the following vocal characteristcs.
Expressive segments
Gemini Caption
Pitch: High female pitch...
Speed: Accelerates with panic...
Emotion: Escalating panic...
......
[Task 1]
3. Instruction Generation
Prompt
Please generate style instructions to guide a Text-to-Speech (TTS) system strictly based on the following voice style description...
Generate instructions
Instructions
[Task 2]
Descriptive-Style Directive
[Task 3]
Role-Play
4. Evaluation
Guidelines
Your task is to judge whether the audio matches the described style ...
Style instruction
TTS Synthesis
Eval
Human
True
False
Gemini
True
False
Further evaluation...

Figure 2: Overview of the benchmark construction and evaluation. We perform careful data cleaning and filtering to select audio with high expressiveness. Then we prompt Gemini to generate detailed, free-form descriptions for each acoustic feature, this caption is also used as the APS task's instruction. Then, we prompt GPT-4o to generate diverse instructions to create diverse DSD and RP instructions. Finally, after consistency verification between human and Gemini as judges, we further perform holistic evaluation on current TTS systems.

tional expressions without bias toward emotional polarity. Finally, we collect 2,000 segments as reference audio. Data statistics can be seen in Tab. 2.

Source	# items		Duration (h)	
	EN	ZH	EN	ZH
NCSSD	183	500	0.47	0.93
Collected	817	500	2.57	1.61
Overall	1,000	1,000	3.04	2.54

Table 2: Statistics of reference audio

3.2.2 Speech Caption

Compared to using predefined categorical tags, we argue that continuous natural language provides a more precise and nuanced description of the relevant speech features. To this end, we leverage the strong spoken language understanding (SLU) capabilities of Gemini [6] to generate a natural language description for each reference audio sample. The prompting strategy used for caption generation is illustrated in Fig. 3. This caption also serves as the instruction for the APS task. In particular, we emphasize the importance of capturing *dynamic changes* in vocal attributes. For example, a speaker might drop to a whisper when starting a gossip; or gradually escalate in emotional intensity, leading to increased volume and even shouting by the end. These dynamic transitions are common in natural speech but have often been overlooked in prior free-form style descriptions. We believe that modeling such temporal transitions is a crucial aspect of evaluating controllability in TTS systems.

3.2.3 Instruction Generation

Based on the generated captions, we use ChatGPT [7] to produce natural language style instructions. For the Descriptive-Style Directive (DSD) task, we randomly drop certain features from the instruction to simulate incomplete or underspecified input. The omitted features are considered unconstrained, allowing the TTS system to generate any plausible values for those attributes. For the Role-play (RP) task, we adopt a chain-of-thought (CoT) prompting strategy, guiding GPT to infer the speaker's role or scenario by reasoning through the twelve defined acoustic features. This approach encourages the model to map low-level acoustic cues to high-level social or contextual roles.

3.3 Metrics

We design an instruction-following metric (Zhou et al., 2024) to evaluate whether the synthesized speech aligns with the given instruction. A brief overview of the evaluation criteria is shown in Tab. 3, while detailed scoring guidelines are provided in App. C. For each subset, the final score is computed as the macro-average of the instruction-following scores across all 1,000 items.

[6] gemini-2.5-pro-preview-05-06

[7] gpt-4o-2024-08-06

Score	**Criteria**
`true`	The sample's primary style attributes (e.g., gender, pitch, rate, emotion) align with the prompt, without conflict.
`false`	At least one key style attribute clearly conflicts with the prompt, or the overall style deviates from the prompt.

Table 3: Scoring Criteria

Accuracy	**APS**	**DSD**	**RP**	**Avg.**
EN	86%	78%	66%	76.7%
ZH	88%	80%	76%	81.3%
Avg.	87%	79%	71%	79.0%

Table 4: Consistency between human majority vote and Gemini.

TTS	**APS**	**DSD**	**RP**	**Avg.**
reference_audio	96.2	89.4	67.2	84.3
Closed-sourced				
gpt-4o-mini-tts	76.4	74.3	**54.8**	68.5
hume*	**83.0**	**75.3**	54.3	**71.1**
Open-sourced				
VoxInstruct	54.9	57.0	39.3	50.4
Parler-TTS-mini	63.4	48.7	28.6	46.9
Parler-TTS-large	60.0	45.9	31.2	45.7
PromptTTS	64.3	47.2	31.4	47.6
PromptStyle	57.4	46.4	30.9	38.2

Table 5: Performance of instruction-following TTS Systems on English (EN). Best results are **bold** and second-best are underlined. * denotes that few cases are missing due to model constraints such as max instruction length or safety settings.

4 Evaluation

4.1 Consistency

To assess whether Gemini can serve as a reliable substitute for human evaluation, we first measure the consistency between human judgments and Gemini's judgments. For each of the three instruction types across two languages, we randomly selected 50 reference audio samples for human annotation. Among them, 25 samples are matched, meaning the instruction was originally derived from the corresponding reference audio. The remaining 25 are mismatched pairs, created by randomly assigning a non-corresponding instruction from the dataset to each reference audio. This setup allows us to assess Gemini's ability to correctly reject negative cases. Note, however, that due to the many-to-many (Ji et al., 2024) nature of speech and description, it is still possible for a mismatched instruction to partially align with the given audio.

We recruit three human annotators (graduate students, all with TOEFL scores above 100) and compensate them at a rate of 50 RMB/hour. They follow the same evaluation guidelines as Gemini. Detailed annotation guidance and screenshots are illustrated in App. C. We take the majority vote among the three annotators as the final human judgment, and compare it with Gemini's predictions. The agreement results are summarized in Tab. 4. Notably, DSD and RP instructions are generated without feeding in the reference audio, so they may deviate from the original audio. Furthermore, because RP prompts are inherently more subjective, human–Gemini agreement tends to decline.

4.2 TTS Systems

We select models that support free-style description as input, without requiring a prompt speech sample. For closed-sourced models, we evaluate gpt-4o-mini-tts and Hume. ElevenLabs is only partially evaluated due to subscription limitations, so we exclude it from evaluation. Since we must specify a voice for gpt-4o-mini-tts, we randomly choose from the provided female/male voice list according to the caption result. For open-sourced TTS, we select Parler-TTS (Lyth and King, 2024), VoxInstruct (Zhou et al., 2024), and the reproducible variants (Ji et al., 2024) of PromptTTS (Guo et al., 2022) and PromptStyle (Liu et al., 2023) for evaluation. Each evaluation session containing all three tasks costs approximately $12.8 for English and $11.9 for Chinese using Gemini-as-a-judge.

4.3 Results and Analysis

4.3.1 Performance on EN Subset

Overall, the closed-sourced commercial models significantly outperform the open-sourced ones. In particular, hume achieves an impressive average score of 71.1. Both hume and gpt-4o-mini-tts handle emotional nuance with remarkable accuracy. However, gpt-4o-mini-tts's need for forced voice selection (we aligned male/female choices) some-

TTS	APS	DSD	RP	Avg.
reference_audio	90.9	86.7	69.8	88.7
Closed-sourced				
gpt-4o-mini-tts	54.9	52.3	46.0	51.1
Open-sourced				
VoxInstruct	47.5	52.3	42.6	47.5

Table 6: Performance of instruction-following TTS Systems on Chinese (ZH).

times conflicts with the natural instructions, which can degrade its performance on timbre-related attributes—such as texture, age, and pitch—that depend on physiological voice quality. Among the open-source models, VoxInstruct stands out: it clearly outperforms its peers on the DSD and RP tasks. Yet in the APS task, its inability to process longer inputs often leads to outputs that are semantically and acoustically undistinguishable. Parler-TTS-large and Parler-TTS-mini show no significant performance gap—and on certain APS and DSD cases, the mini version even edges out the large. PromptTTS and PromptStyle, by contrast, struggle to generate expressive speech, yielding rather flat, unremarkable samples.

Nonetheless, across all TTS systems, there remains a substantial expressiveness gap between synthesized output and the reference audio. Bridging this divide to reach truly free-form, human-level control and naturalness in TTS remains an open challenge.

4.3.2 Performance on ZH Subset

In the ZH subset, gpt-4o-mini-tts slightly outperforms VoxInstruct (+3.6), as shown in Tab. 6. Though its timbre can sometimes conflict with the specified instructions, it handles emotional expression very effectively. VoxInstruct exhibits instability when given APS instructions, failing to follow the script, resulting in an undistinguishable voice style. This issue is largely mitigated for DSD and RP tasks, likely because those instructions more closely resemble its training data. Meanwhile, it performs relatively well on “news anchor” prompts but struggles with more expressive directives. Notably, whenever VoxInstruct does stick to the script, its Mandarin prosody sounds appreciably more natural than gpt-4o-mini-tts’s. Moreover, performance on the ZH subset lags significantly behind that on the EN subset, highlighting a substantial disparity in TTS capabilities across languages.

4.4 Case Studies

In this section, we select some representative cases for analysis (Tab. 7).

First, **modern TTS models still struggle to reproduce the paralinguistic *sound events* that frequently occur in human speech**, such as sighs, sudden bursts of laughter, screams, etc. In our case study, only gpt-4o-mini-tts is able to generate laughter. Yet these vocal events are essential for conveying emotion and maintaining a natural speech flow; a powerful, controllable TTS system should be able to capture and synthesize them.

Few models are capable of extreme emotional expressions and rapid affective shifts. In our evaluation, gpt-4o-mini-tts clearly stands out: it can produce controlled shouting and other heightened vocalizations on demand. Meanwhile, VoxInstruct and gpt-4o-mini-tts are also initially capable of handling some transitions—such as moving from calm speech to excited.

Interestingly, in certain timbre-focused cases, **some open-sourced systems actually can produce surprisingly impressive results**. Take the “child voice” scenario, for example: hume is blocked from generating child-like timbres by its safety filters, and gpt-4o-mini-tts cannot stray from its fixed voice settings, which largely limit its prosodic generality. Yet several open-source TTS models—despite weaker overall emotional control—deliver astonishingly authentic, youthful vocal qualities. And Parler-TTS-Large successfully synthesizes the voice of the elderly. This suggests that timbre flexibility and emotional expressiveness remain orthogonal capabilities, and that future TTS research should aim to unify them rather than treat them separately.

Finally, it’s worth noting that **no current system can truly generate a “singing” effect**. Instructing a TTS model to “speak as if singing” requires integrating prosody, melody, emotion, timbre, and rhythm in a cohesive way. This multidimensional coordination sets a much higher bar for naturalness and expressiveness in controllable TTS—and represents a key direction for future research.

5 Conclusions

In this paper, we carefully construct a hirerachical benchmark for instruction-following TTS.

Instruction	Performance
Acoustic-Parameter Specification	
Speed: Rapid pace initially, slightly slowing. Volume: Energetic and relatively loud, decreasing slightly. Emotion: Excited and emphatic, ending with a sigh suggesting weariness.	**NO** existing models successfully ‘sigh’; hume and gpt-4o-mini-tts sound excited.
Volume: Shouting, very loud. Texture: Tense, somewhat strained. Emotion: Intense panic and fear.	gpt-4o-mini-tts shows signs of ‘shouting’; hume bears anger.
Descriptive-Style Directive	
Begin with an artificially high-pitched, boisterous laugh that carries a playful tone, then smoothly transition to a more deliberate pace with a standard conversational volume, subtly lowering the pitch afterward to deliver the rest with a slightly put-upon nasal quality.	gpt-4o-mini-tts successfully laughs out.
Infuse your performance with an outgoing personality, ensuring a high child pitch is woven through a swift, energetic delivery.	VoxInstruct, Parler-TTS-large generates voice like a child; gpt-4o-mini-tts outputs energetic delivery.
Role-Play	
Use an expressive and somewhat theatrical tone, like an elderly British female storyteller sharing a funny story at a family gathering. Start with a quick pace and clear articulation, then slow down, slightly fluctuating in pitch to emphasize key parts with a quirky and slightly bossy texture.	Parler-TTS-large generates a trembling voice of an elderly; VoxInstruct delivers a middle-aged to elderly female voice.
Create an effect that keeps listeners focused: imagine a scenario where someone is shouting in panic and agony, their words blurred and barely comprehensible, with piercing screams that demand immediate attention.	**NO** models scream; gpt-4o-mini-tts speaks as if it is in pain.
Share the message with the energy of a young adult cartoon character, starting with a clear and calm voice that quickly rises to a dynamic, emotionally impulsive pitch.	VoxInstruct clearly raises its voice; gpt-4o-mini-tts shows a slight rise.
Infuse your tone with the brightness of a stage performer in a whimsical play, keeping your voice clear, lighthearted, and effortlessly melodic.	**NO** models generate melodic voice.

Table 7: Case studies. Models not mentioned indicate a lack of significant expressiveness. For APS instructions, some parameters are omitted due to length constraints.

We meticulously design a three-tier evaluation task—spanning the low-level Acoustic-Parameter Specification (APS) task, the mid-level Descriptive-Style Directive (DSD) task, and the high-level Role-Play (RP) task—to comprehensively measure current TTS systems’ ability to follow complex natural-language descriptions of acoustic features. Our results reveal that existing models still struggle with fine-grained paralinguistic control, and expose significant performance gaps both between closed- and open-source systems and across different languages. Moreover, our case studies highlight major deficiencies in reproducing natural vocal events, managing extreme emotional transitions, and synthesizing character-specific timbres—capabilities that are crucial for advancing TTS toward truly human-like expressiveness. We hope this benchmark will catalyze further progress in developing more controllable and expressive speech synthesis.

Limitations

We acknowledge that our work may have the following limitations: 1) Subjectivity and evaluation cost. Some of our tasks, particularly for the Role-Play (RP), are inherently subjective. Inter-annotator agreement among human raters is relatively low compared to APS and DSD tasks, which introduces noise when using automated evaluators like Gemini. Moreover, continuously conducting large-scale evaluations using Gemini is cost-intensive. In future work, we plan to develop a more accurate and cost-efficient evaluator to enact iterative evaluation. 2) Data imbalance. Since our dataset is constructed in a bottom-up fashion from specific acoustic events and style directives, certain classes (e.g., particular emotions or role archetypes) are underrepresented. This imbalance could bias model performance and evaluation. We will expand the benchmark to include a broader, more evenly distributed set of style categories.

Ethical Considerations

Owing to our large-scale automated pipeline, we are unable to manually review every text–instruction pair. As a result, it may consist of some inappropriate content. Please note that any content from the reference audio and the synthesized audio does NOT reflect the authors' views or endorsements. Additionally, this dataset and benchmark are intended solely for academic and research use.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *ArXiv*, abs/2406.02430.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *ArXiv*, abs/2406.13340.

BELLEGroup. 2023. Belle: Be everyone's large language model engine. `https://github.com/LianjiaTech/BELLE`.

James Betker. 2023. Better speech synthesis through scaling. *ArXiv*, abs/2305.07243.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *ArXiv*, abs/2410.06885.

Anne Cutler, Delphine Dahan, and Wilma Van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and speech*, 40(2):141–201.

Jordan Darefsky, Ge Zhu, and Zhiyao Duan. 2024. Parakeet.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. 2025. Scaling rich style-prompted text-to-speech datasets. *ArXiv*, abs/2503.04713.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *ArXiv*, abs/2407.05407.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jing-Ru Zhou. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *ArXiv*, abs/2412.10117.

Hao-Han Guo, Kun Liu, Fei-Yu Shen, Yi-Chen Wu, Fenglong Xie, Kun Xie, and Kai-Tuo Xu. 2024. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *ArXiv*, abs/2409.03283.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xuejiao Tan. 2022. Prompttts: Controllable text-to-speech with text descriptions. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2023. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Siqi Zheng, Qian Chen, Wen Wang, Ziyue Jiang, Hai Huang, Xize Cheng, Rongjie Huang, and Zhou Zhao. 2024. Controlspeech: Towards simultaneous zero-shot speaker cloning and zero-shot language style control with decoupled codec. *ArXiv*, abs/2406.01205.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 2024. Speechcraft: A fine-grained expressive speech dataset with natural language description. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. 2021. Expressive text-to-speech using style tag. In *Interspeech*.

Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiang-Yang Li, Sheng Zhao, Tao Qin, and Jiang Bian. 2023. Prompttts 2: Describing and generating voices with text prompt. *ArXiv*, abs/2309.02285.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *Preprint*, arXiv:2411.01156.

Guanghou Liu, Yongmao Zhang, Yinjiao Lei, Yunlin Chen, Rui Wang, Zhifei Li, and Linfu Xie. 2023. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. In *Interspeech*.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. 2024. Generative expressive conversational speech synthesis. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Daniel Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *ArXiv*, abs/2402.01912.

Gallil Maimon, Amit Roth, and Yossi Adi. 2024. A suite for acoustic language model evaluation. *ArXiv*, abs/2409.07437.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *ArXiv*, abs/2006.04558.

Apoorv Vyas, Bowen Shi, Matt Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, W.K.F. Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Tunde Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, and 5 others. 2023. Audiobox: Unified audio generation with natural language prompts. *ArXiv*, abs/2312.15821.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn Schuller. 2022. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10745–10759.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2024a. Audiobench: A universal benchmark for audio large language models. *ArXiv*, abs/2406.16020.

Chengyi Wang, Sanyuan Chen, Yu Wu, Zi-Hua Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, and 6 others. 2025. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *Preprint*, arXiv:2503.01710.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024b. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *ArXiv*, abs/2409.00750.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Guangzhi Lei, Chao Weng, Helen M. Meng, and Dong Yu. 2023. 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.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, Fan Yu, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting.

Zhen Ye, Xinfa Zhu, Chi min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yi-Ting Guo, and Wei Xue. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *ArXiv*, abs/2502.04128.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shunwei Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. 2024. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. *Proceedings of the 32nd ACM International Conference on Multimedia*.

A Speech Caption

Fig. 3 illustrates our prompt for Gemini to generate a detailed, natural language description for each speech segments.

B Instruction Generation

In this section, we introduce our prompting strategies for generating style instructions. In order to increase diversity and accuracy, we employ: 1) random dropouts, 2) multiple prompts, 3) generate more instructions at one time, and 4) utilize the Chain-of-Thought (CoT) strategy.

To generate the DSD prompt, we design 3 weight settings to randomly drop out features, as shown in Fig. 8. When instructing GPT-4o to generate instructions, we select a weight setting and provide last features to the model. Prompt for generating DSD can be seen in Fig. 4. Also, we design 3 prompts for generating RP instructions, as shown in Fig. 5 , Fig. 6, and Fig. 7.

	Set 1	Set 2	Set 3
probs	0.5	0.25	0.25
gender	1.0	0.5	0.5
pitch	1.0	0.5	0.5
speed	1.0	0.5	0.5
volume	1.0	0.5	0.5
age	1.0	0.8	0.5
clarity	1.0	0.5	0.5
fluency	1.0	0.5	0.5
accent	1.0	0.8	0.5
texture	1.0	0.8	0.5
emotion	1.0	1.0	0.8
tone	1.0	1.0	0.8
personality	1.0	1.0	0.8

Table 8: Weights

C Scoring Guidelines

Fig. 8 illustrates our scoring guidelines for Gemini, and the screenshot of human annotation can be seen in Fig. 9.

```
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.
```

Figure 3: Speech caption prompt for Gemini.

~~~~~
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.
~~~~~

Figure 4: Prompts for DSD instruction generation.

~~~~~
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

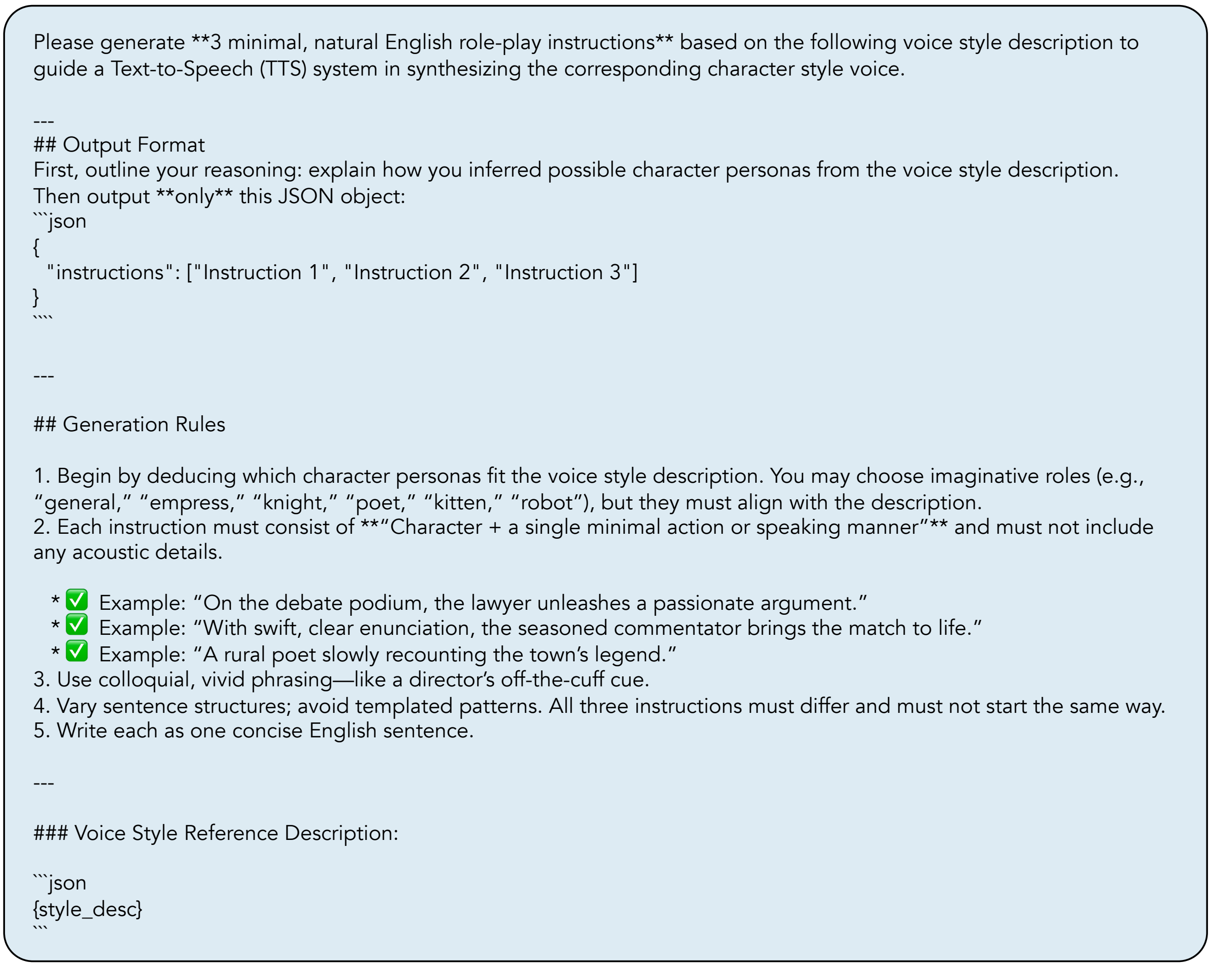
```json
{style_desc}
```

Please return **only** the JSON object—no extra text.
~~~~~

Figure 5: Prompts for RP instruction generation.

~~~
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)

* ❌ "Pretend you are a cold assassin."
* ❌ "Act like a weary teacher."
* ❌ "Speak in a calm voice with moderate pace."

---

### Voice Style Reference Description

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

Figure 6: Prompts for RP instruction generation.



~~~~~text
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.

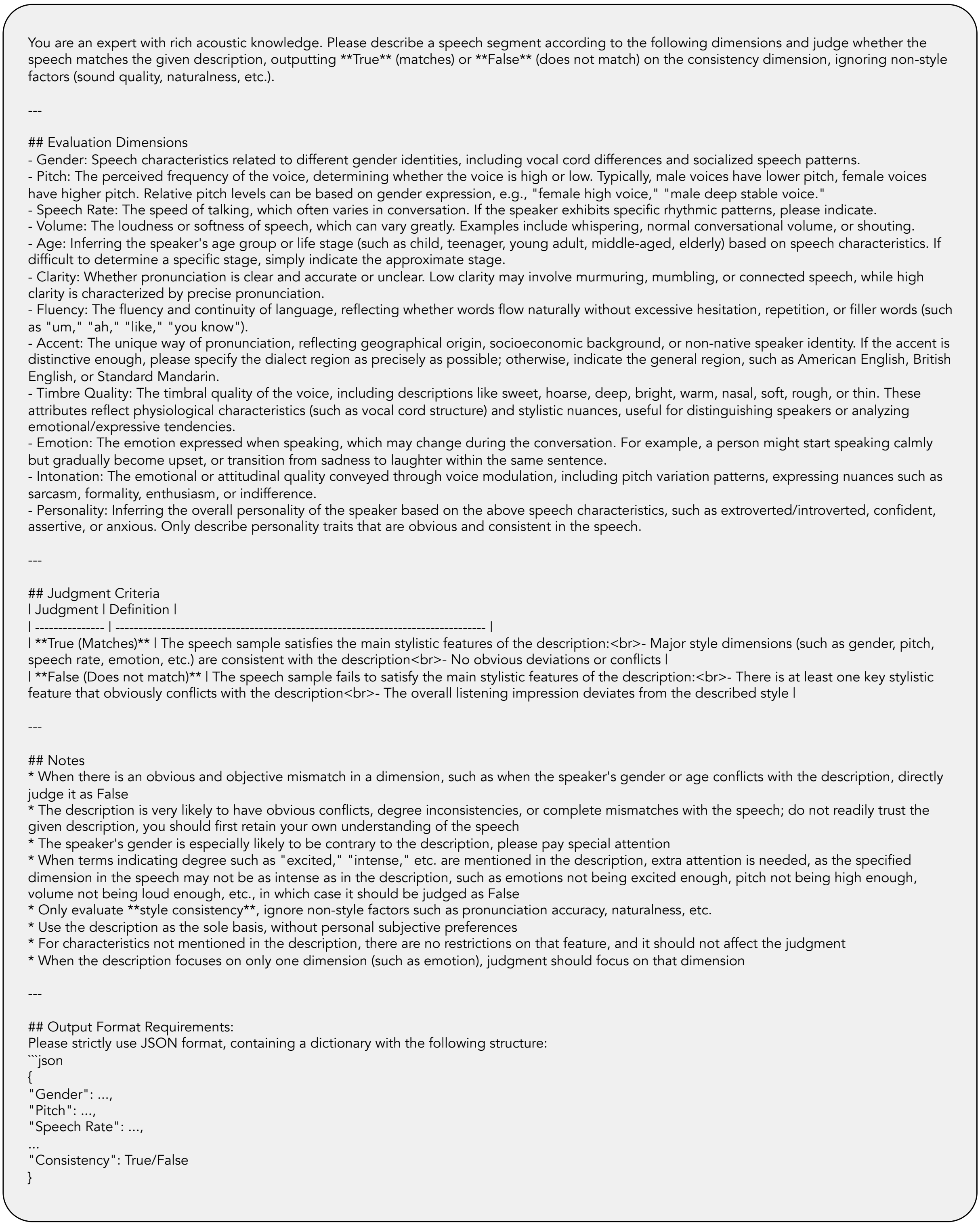
  * ✅ Example: "On the debate podium, the lawyer unleashes a passionate argument."
  * ✅ Example: "With swift, clear enunciation, the seasoned commentator brings the match to life."
  * ✅ Example: "A rural poet slowly recounting the town's legend."
3. Use colloquial, vivid phrasing—like a director's off-the-cuff 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:

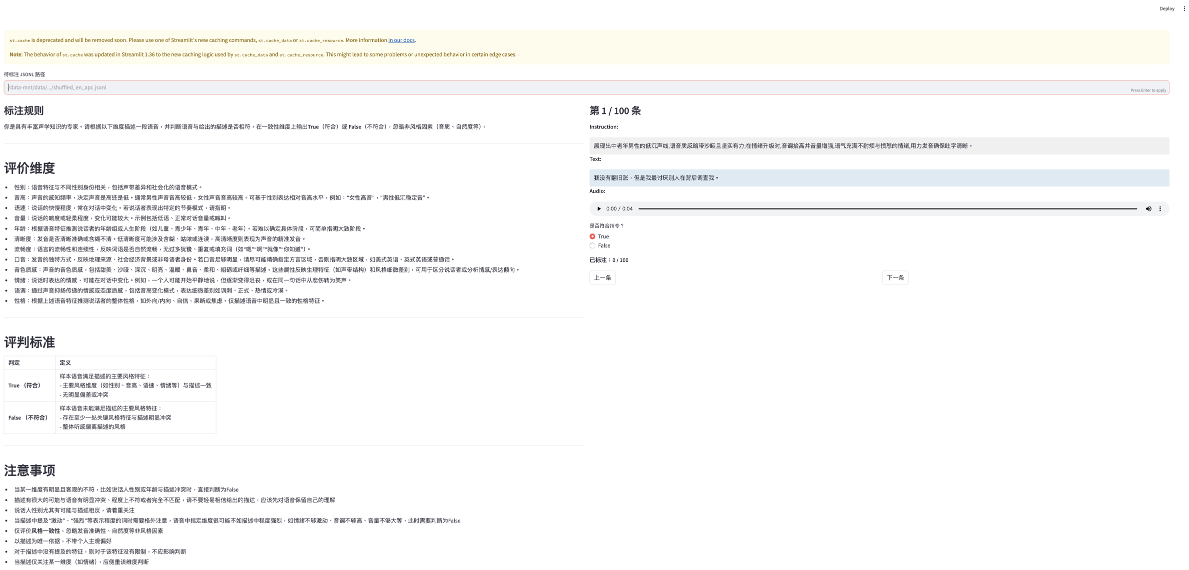
```json
{style_desc}
```
~~~~~

Figure 7: Prompts for RP instruction generation.



~~~
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
| Judgment | Definition |
| --------------- | ------------------------------------------------------------------------------------ |
| **True (Matches)** | The speech sample satisfies the main stylistic features of the description:<br>- Major style dimensions (such as gender, pitch, speech rate, emotion, etc.) are consistent with the description<br>- No obvious deviations or conflicts |
| **False (Does not match)** | The speech sample fails to satisfy the main stylistic features of the description:<br>- There is at least one key stylistic feature that obviously conflicts with the description<br>- 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
}
~~~

Figure 8: Prompt for Gemini-as-a-judge (translated ver.).



st.cache is deprecated and will be removed soon. Please use one of Streamlit's new caching commands, st.cache_data or st.cache_resource. More information in our docs.

Note: The behavior of st.cache was updated in Streamlit 1.36 to the new caching logic used by st.cache_data and st.cache_resource. This might lead to some problems or unexpected behavior in certain edge cases.

待标注 JSONL 路径

/data-mnt/data/.../shuffled_en_apc.jsonl

标注规则

你是具有丰富声学知识的专家。请根据以下维度描述一段语音，并判断语音与给出的描述是否相符，在一致性维度上输出True（符合）或 False（不符合），忽略非风格因素（音质、自然度等）。

评价维度

- 性别：语音特征与不同性别身份相关，包括声带差异和社会化的语音模式。
- 音高：声音的感知频率，决定声音是高还是低。通常男性声音音高较低，女性声音音高较高。可基于性别表达相对音高水平，例如："女性高音"，"男性低沉稳定音"。
- 语速：说话的快慢程度，常在对话中变化。若说话者表现出特定的节奏模式，请指明。
- 音量：说话的响度或轻柔程度，变化可能较大。示例包括低语、正常对话音量或喊叫。
- 年龄：根据语音特征推测说话者的年龄组或人生阶段（如儿童、青少年、青年、中年、老年）。若难以确定具体阶段，可简单指明大致阶段。
- 清晰度：发音是否清晰准确或含糊不清。低清晰度可能涉及含糊、咕哝或连读，高清晰度则表现为声音的精准发音。
- 流畅度：语言的流畅性和连续性，反映词语是否自然流畅，无过多犹豫、重复或填充词（如"嗯""啊""就像""你知道"）。
- 口音：发音的独特方式，反映地理来源、社会经济背景或非母语者身份。若口音足够明显，请尽可能精确指定方言区域，否则指明大致区域，如美式英语、英式英语或普通话。
- 音色质感：声音的音色质感，包括甜美、沙哑、深沉、明亮、温暖、鼻音、柔和、粗砺或纤细等描述。这些属性反映生理特征（如声带结构）和风格细微差别，可用于区分说话者或分析情感/表达倾向。
- 情绪：说话时表达的情感，可能在对话中变化。例如，一个人可能开始平静地说，但逐渐变得沮丧，或在同一句话中从悲伤转为笑声。
- 语调：通过声音抑扬传递的情感或态度质感，包括音高变化模式，表达细微差别如讽刺、正式、热情或冷漠。
- 性格：根据上述语音特征推测说话者的整体性格，如外向/内向、自信、果断或焦虑。仅描述语音中明显且一致的性格特征。

评判标准

判定	定义
True （符合）	样本语音满足描述的主要风格特征： - 主要风格维度（如性别、音高、语速、情绪等）与描述一致 - 无明显偏差或冲突
False （不符合）	样本语音未能满足描述的主要风格特征： - 存在至少一处关键风格特征与描述明显冲突 - 整体听感偏离描述的风格

注意事项

- 当某一维度有明显且客观的不符，比如说话人性别或年龄与描述冲突时，直接判断为False
- 描述有很大的可能与语音有明显冲突、程度上不符或者完全不匹配，请不要轻易相信给出的描述，应该先对语音保留自己的理解
- 说话人性别尤其有可能与描述相反，请着重关注
- 当描述中提及"激动"、"强烈"等表示程度的词时需要格外注意，语音中指定维度很可能不如描述中程度强烈，如情绪不够激动、音调不够高、音量不够大等，此时需要判断为False
- 仅评价**风格一致性**，忽略发音准确性、自然度等非风格因素
- 以描述为唯一依据，不带个人主观偏好
- 对于描述中没有提及的特征，则对于该特征没有限制，不应影响判断
- 当描述仅关注某一维度（如情绪），应侧重该维度判断

第 1 / 100 条

Instruction:

展现出中老年男性的低沉声线,语音质感略带沙哑且坚实有力;在情绪升级时,音调抬高并音量增强,语气充满不耐烦与愤怒的情绪,用力发音确保吐字清晰。

Text:

我没有翻旧账，但是我最讨厌别人在背后调查我。

Audio:

是否符合指令？

● True
○ False

已标注：0 / 100

Figure 9: Screenshot for human annotation.