EXPRESSTTS-BENCH: BENCHMARKING THE SCENARIO-BASED EXPRESSIVENESS OF TEXT-TO-SPEECH MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent Text-to-Speech (TTS) systems have achieved substantial progress in sound quality and integrity, yet their ability to generate natural and appropriate highly expressive audio based on a given scenario is still very limited. Meanwhile, the existing evaluation of expressiveness still relies on costly and inefficient human evaluation. Moreover, existing expressiveness evaluation typically overlooks the scenario and textual context that fundamentally shape how expressiveness should be perceived. To this end, we introduce ExpressTTS-Bench, a fully automated benchmark for evaluating the scenario-based expressiveness of TTS models. ExpressTTS-Bench constructs diverse test data through a hierarchical scenario tree covering five representative expression modes and four task formats, resulting in $8,000$ test samples across English and Chinese. To enable fine-grained and scalable evaluation, we decompose expressiveness into nine sub-dimensions spanning emotion, prosody, and paralanguage, and introduce a LALM-as-a-Judge framework with automatically generated, expression-mode-specific scoring rubrics. We evaluate six representative open-source and closed-source TTS models and reveal substantial differences in scenario-based expressiveness across models, languages, instruction formats, expression modes, and sub-dimensions. Consistency analyses and human evaluation further support the reliability of ExpressTTS-Bench. ExpressTTS-Bench provides a scalable framework for systematically evaluating the scenario-conditioned expressive capabilities of modern TTS systems.

1 INTRODUCTION

Recent speech models have demonstrated outstanding audio generation capabilities, among which Text-to-Speech (TTS) models are able to generate high-quality audio based on specific context (Zhang et al., 2025a;b; Hu et al., 2026; Gong et al., 2026; Zhou et al., 2026). However, although current TTS systems have already been capable of generating complete and usable audio, the generated audio often sounds mechanical and unnatural, especially lacking sufficient expressive tension in different scenarios (Manku et al., 2026; Liao et al., 2026; Hu et al., 2026).

Meanwhile, conducting effective evaluations is also crucial for understanding and further improving the model's performance. Some traditional works evaluate the TTS systems by designing some automated metrics, such as speaker-similarity score (SIM), word error rate (WER), and so on (Anastassiou et al., 2024; Xu et al., 2025; Hu et al., 2026), which are used for some basic acoustic quality and accuracy evaluation. Additionally, some subjective metrics such as Mean Opinion Score (MOS) are utilized to evaluate some comprehensive or subjective dimensions (Viswanathan & Viswanathan, 2005; Zhou et al., 2026). However, this type of method mainly relies on human annotation, which is inefficient, expensive, and difficult to reproduce and scale. Considering the limitations of human evaluation, Manku et al. (2026) are the first to deeply investigate the model-as-a-judge paradigm of applying Large Audio Language Model (LALM) as a reward model to audio evaluation.

As an important subjective attribute, the current assessment of expressiveness mainly relies on subjective MOS tests. On the one hand, objective acoustic features are difficult to reflect subtle performance variations and specific contextual situations; on the other hand, human annotation greatly limits the efficiency and scalability of expressiveness evaluation (Manku et al., 2026). Lin et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2025) further try to objectify expressiveness by aligning with human preference. However, this framework still relies on human-annotated data for fine-tuning. More importantly, in reality, the perception and evaluation of expressiveness are closely tied to the specific scenario and text, while current evaluation framework only focuses on the audio itself, ignoring the crucial scene and textual information. Overall, there is still a lack of a dedicated, systematic, and efficient evaluation framework for assessing the expressiveness of TTS models.

In this work, we propose ExpressTTS-Bench, a new automatic benchmark specially designed to evaluate the scenario-based expressiveness of TTS models, which includes test data construction based on diverse scenarios, exclusive rubrics generation for each expression mode and sub-dimension, and LALM-as-a-Judge expressiveness assessment. ExpressTTS-Bench introduces a hierarchical tree structure for automatic scenario-based data construction. Following Huang et al. (2025), we also design different tasks in the construction of test samples, including *Direct*, *Attribute List*, *Natural Language*, and *Scenario Performance Guidance*. We finally automatically construct a total of $8,000$ test samples. Through this paradigm, we can also expand the test set on more new expression modes.

Additionally, we utilize LALM-as-a-Judge paradigm and break down expressiveness into 9 specific sub-dimensions. Because different expression modes differ and the evaluation targets of different sub-dimensions are different, we use exclusive evaluation rubrics for each expression mode and each sub-dimension. Considering that it is very difficult to manually design evaluation rubrics for a large number of combinations, we also propose a paradigm that generates corresponding exclusive rubrics for each different expression mode and sub-dimension based on a small number of reference audios. It not only achieves efficient automated multi-scenario, multi-dimensional evaluation, but also demonstrates good consistency with human evaluation.

We conduct the experiments on 6 different TTS models. Our results indicate that current TTS models still have significant shortcomings in generating highly expressive, natural, and contextually appropriate audio based on specific scenarios and text, and there are also notable capability gaps among different models. We also analyze and verify the consistency, validity, and stability of our evaluation framework. All results establish ExpressTTS-Bench as a valuable benchmark for evaluating the expressiveness of TTS models.

Our contributions are as follows:

- We propose **ExpressTTS-Bench**, a **automatic** and **scalable** benchmark designed for evaluating the **scenario-based expressiveness** of TTS models, which explicitly incorporates scenario and textual context into the evaluation process, enabling a more comprehensive expressiveness assessment.
- We develop an automated and scalable scenario-based test data construction pipeline, which introduces a scalable hierarchical tree structure to construct diverse scenarios. We also introduce a fine-grained **LALM-as-a-Judge** evaluation framework for scenario-based expressiveness assessment from nine sub-dimensions. For each different expression mode and sub-dimension, we **automatically derive exclusive evaluation rubrics**, enabling efficient and scalable multi-scenario and multi-dimensional assessment while maintaining strong agreement with human judgments.
- We conduct extensive experiments on six representative TTS models using ExpressTTS-Bench, revealing substantial differences in scenario-conditioned expressive capabilities across existing systems and highlighting their remaining limitations in generating natural, highly expressive, and contextually appropriate audio.

2 Related Work

2.1 Expressive and Controllable Text-to-Speech Generation

TTS models have progressed from learning latent speaking styles to controlling speech through instruction descriptions. Early works introduce reference-based prosody transfer and global style tokens to capture variation beyond linguistic content (Skerry-Ryan et al., 2018; Wang et al., 2018), while FastSpeech 2 further explicitly model pitch, energy, and duration (Ren et al., 2020). Subsequent generative approaches advance style diversity and zero-shot synthesis through style diffu-

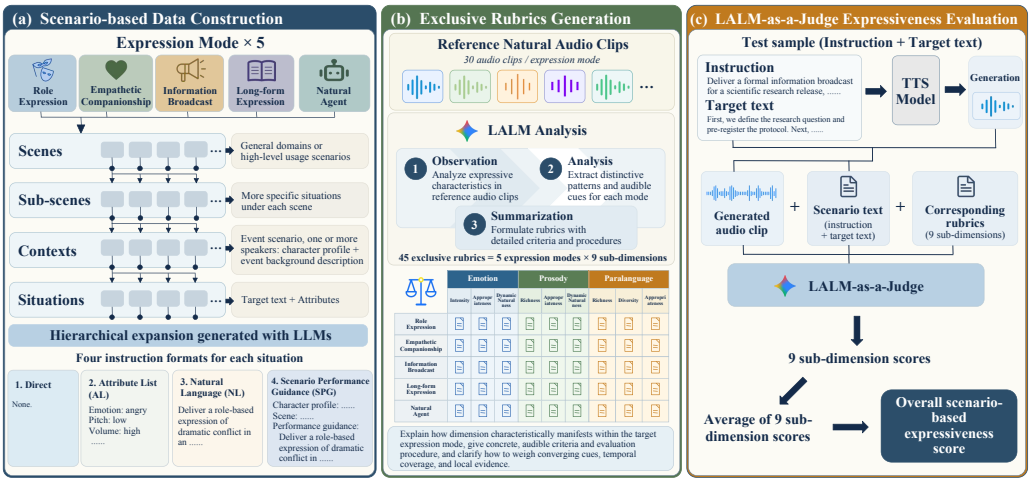


Figure 1: Overview of ExpressTTS-Bench. It includes scenario-based test data construction (*Left*), exclusive rubrics generation for all expression modes and all sub-dimensions (*Center*), and LALM-as-a-Judge automatic expressiveness evaluation (*Right*).

sion (Li et al., 2023), flow-based speech infilling (Le et al., 2023), and latent or factorized diffusion (Shen et al., 2024; Ju et al., 2024). In parallel, some works make natural language an interface for specifying speaking styles and voice characteristics (Guo et al., 2023; Yang et al., 2024; Leng et al., 2024). Although these advances have broaden the range of realizations available for a given text, generative controllability alone does not establish whether a realization is highly expressive and appropriate for its context. Thus, ExpressTTS-Bench focuses on the joint assessment of expressive richness, contextual appropriateness, and temporal naturalness across emotion, prosody, and paralanguage, which organizes test cases around a hierarchical scenario tree spanning five expression modes, four instruction formats, and two languages.

2.2 Speech Evaluation

Speech evaluation spans signal fidelity, intelligibility, perceptual quality, expressive behavior, and so on. Classical reference-based metrics estimate perceived quality under transmission distortions and intelligibility under noise, respectively (Rix et al., 2001; Taal et al., 2011). Modern TTS evaluation also commonly reports transcription word error rate (WER) and speaker similarity (SIM) (Manku et al., 2026). Human listening tests based on MOS remain central to speech assessment, such as large-scale naturalness annotations SOMOS (Maniati et al., 2022), pairwise preference judgments (Zhang et al., 2026), and so on. To scale evaluation, some works try to predict human ratings (Lo et al., 2019; Saeki et al., 2022). In addition, building on LLM-as-a-Judge (Zheng et al., 2023), large audio-language models (LALMs) further enable evaluation through audio understanding and judgment (Wang et al., 2025; Zhang et al., 2026; Manku et al., 2026; Li et al., 2026). For expressiveness evaluation, Lin et al. (2025) try to align emotion, prosody, and spontaneity assessments with human preferences (Lin et al., 2025). However, this framework still relies heavily on human-annotated data for fine-tuning. ExpressTTS-Bench achieves an automated evaluation pipeline from test data construction to the generation of expression-mode-specific evaluation rubrics, and then to LALM-as-a-Judge evaluation. ExpressTTS-Bench not only provides an effective framework for scenario-based expressiveness evaluation, but also has good scalability.

3 ExpressTTS-Bench

We illustrate the overall framework of ExpressTTS-Bench in Figure 1, which includes a fully automated pipeline covering multi-scenario test data construction, exclusive rubrics generation for each expression mode and sub-dimension, and model expressiveness evaluation. In this section, we introduce the ExpressTTS-Bench in detail.

3.1 SCENARIO-BASED DATA CONSTRUCTION

We emphasize the diversity and representativeness of the data in our ExpressTTS-Bench dataset. To comprehensively and effectively evaluate the expressiveness of TTS models, we propose a scenario-centric framework for constructing test data. Specifically, we perform hierarchical scene expansion through a tree structure, which can progressively expand from a rough expression mode category to a large number of specific situations. The entire construction and generation process is completed based on Gemini-3.6-Flash.

Initially, we select five representative **Expression Modes** for the benchmark: role expression, empathetic companionship, information broadcast, long-form expression, and natural agent. Then, we further expand finer-grained **Scenes** for each expression mode. After generating finer-grained scenes for each expression mode, we further generate finer-grained **Sub-scenes** for each scene. Furthermore, we generate **Contexts** for each sub-scene. Each context here corresponds to a specific event scenario, which may contain one or more speakers. Each context contains one or more items depending on the number of speakers, with each item corresponding to a specific speaker's profile as well as event-related description from their corresponding perspective. After this, each item can be expanded one-to-one into a corresponding **Situation**, which contains target text and specific descriptions of attributes such as emotion, volume, pitch, and so on. During this process, to control the granularity of generation, we prepare some "Expression Mode - Scene - Sub-scene" chain seed data as few-shot references. We present some detailed examples in Appendix A.

After all situation end nodes are generated, we further generate the final test sets. Inspired by Huang et al. (2025), we design four different types of tasks based on instruction styles:

1. *Direct*: These samples use empty instructions, requiring the model to directly generate audio without relying on instructions. This is designed to assess the model's ability to understand the context based on the text and generate appropriate, highly expressive audio.
2. *Attribute List* (AL): In this task, we organize different attributes and their corresponding values in lists. Each instruction is a list-style combination of the attributes contained in the situation sample, which contains descriptions of 10 different fields.
3. *Natural Language* (NL): We describe multiple attributes and their values in the situation sample using natural language. The description style of natural language is closer to daily human expression, testing the model's performance on simple unstructured instructions.
4. *Scenario Performance Guidance* (SPG): In this task, we use complex scenario-based description as instruction, which provide detailed performance guidance for the model. Each instruction consists of three fixed parts: *Character profile*, *Scene*, and *Scenario-based performance guidance*. Such complex instructions can test the model's understanding and performance capabilities regarding scenarios and fine-grained guidance.

We show the prompts used to generate the test sets in the different formats and specific examples of test samples in different formats in Appendix B. Overall, we comprehensively evaluate the TTS models' capability to generate highly expressive audio through different types of tasks.

3.2 EXCLUSIVE RUBRICS GENERATION

To more comprehensively evaluate the scenario-based expressiveness of TTS models, we break down expressiveness into nine different sub-dimensions for analysis. These complex and diverse settings, while enabling a more fine-grained evaluation and analysis of model expressiveness, also pose challenges for establishing scoring rubrics. The objects of observation in different sub-dimensions often have many differences, and the evaluation criteria for different expression modes within the same sub-dimension may even contradict each other. For example, depending on the specific context, empathetic companionship often requires stable and supportive emotional expression, whereas role expression often requires dramatic tension and exaggerated contrasts in emotional intensity.

Therefore, we apply exclusive rubrics for each expression mode and each sub-dimension. However, since we set up nine sub-dimensions and five expression modes in total, there will be 45 different combinations in all. In addition, considering the scalability of our test data, the total number of all combinations may further increase. It is very difficult to manually set an evaluation criterion

for each combination, and it may also introduce human errors and biases. Therefore, we design an rubrics generation pipeline that can automatically design corresponding evaluation criteria for different expression modes and sub-dimensions.

We have LALM learn the characteristics of highly expressive audio across different expression modes and different sub-dimensions by sequentially performing observation, analysis, and summarization. First, we collected 30 matching natural audio clips as reference audio for each expression mode. Then, for each expression mode, we use Gemini-3.6-Flash to observe and analyze the characteristics of each audio clip from nine sub-dimensions respectively. Then, for each sub-dimension under each expression mode, LALM further comprehensively summarizes the characteristics exhibited by the reference audios and forms a detailed report. Finally, LALM generates detailed evaluation criteria based on the corresponding reports. The scoring rubric will explain how this dimension characteristically manifests within the target expression mode, while also giving concrete, audible criteria and a step-by-step evaluation procedure and clarifying how an evaluator should weigh converging cues, temporal coverage, and local evidence. Details regarding the generation of scoring rubrics and their content are all in the Appendix F.

3.3 LALM-AS-A-JUDGE EXPRESSIVENESS EVALUATION

Our current dataset contains a total of $8,000$ test samples, which means that each TTS model will ultimately generate $8,000$ audio clips to be evaluated. Evaluating such a large number of samples based on human is very expensive and time-consuming, difficult to reproduce, and hard to guarantee in quality. Thus, following Manku et al. (2026), we utilize the LALM-as-a-Judge paradigm to evaluate the scenario-based expressiveness of the TTS models. Specifically, we utilize Gemini-3.6-Flash to conduct the LALM-as-a-Judge evaluation. We independently score a single audio clip from nine different sub-dimensions, which can be categorized into three main aspects:

- ***Emotion***: Highly expressive audio should convey rich and appropriate emotions. **Emotion Intensity** focuses on the degree of intensity of the emotion expressed in the audio. **Emotion Appropriateness** mainly examines whether the emotion expressed matches the instruction and the scene context. **Emotion Dynamic Naturalness** evaluates how naturally emotional type and intensity are organized over time, including aspects such as continuity, development, local peaks, reductions, and transitions.
- ***Prosody***: Highly expressive audio should also have rich, natural, and reasonable prosodic variation. **Prosody Richness** evaluates the degree of prosodic variation contained in the audio. **Prosody Appropriateness** examines how well the prosodic variations appearing match the character, scene, text, and communicative intent. **Prosody Dynamic Naturalness** assesses temporal coherence and human-like development rather than contextual fit.
- ***Paralanguage***: Paralanguage contains information that words cannot convey, which is very important for the expressiveness of audio. **Paralanguage Richness** represents how much paralinguistic information is present. **Paralanguage Diversity** measures how many types of paralanguage appear. **Paralanguage Appropriateness** mainly focuses on how well the non-lexical vocal events that are actually audible fit the character, scene, target text, and intended communicative function.

Each sub-dimension is scored on a scale of 1 to 10, and then we take the average score of all sub-dimensions as the overall scenario-based expressiveness score on that sample.

We can divide all nine sub-dimensions into two categories: Degree-axis and Quality-axis (Appendix F for details). It should be noted that on the degree-axis sub-dimensions (Table 17 for details), we evaluate based on the reasonable range of the corresponding expression mode, rather than always mapping a score to the same absolute acoustic amplitude.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We conduct our experiments on 6 different representative TTS voice design models. We choose four representative open-source TTS voice design models: Qwen3-TTS-12Hz-1.7B-

Table 1: Overall scenario-based expressiveness scores of eight different sets across six different TTS models. **Bold** indicates the highest score. * indicates that a small number of samples are removed due to the model's maximum length limit.

Model	EN				ZH				Avg. ↑
	Direct ↑	AL ↑	NL ↑	SPG ↑	Direct ↑	AL ↑	NL ↑	SPG ↑	
Elevenlabs eleven-ttv-v3 *	61.52	**75.54**	**75.54**	**74.44**	65.79	77.33	77.60	76.36	73.01
Qwen3-TTS-12Hz-1.7B-VoiceDesign	**66.59**	74.79	74.04	73.90	71.44	**78.61**	**78.37**	**78.31**	**74.51**
Fish Audio voice-design-1	57.02	60.24	60.29	59.77	70.17	73.64	72.96	73.18	65.91
VoxCPM2	60.79	72.49	69.47	72.95	**71.98**	77.69	77.15	76.87	72.42
MOSS-VoiceGenerator	58.19	52.22	61.76	60.17	70.92	68.39	73.12	71.14	64.49
MiMo-Audio-7B-Instruct	57.75	69.44	68.98	66.67	67.35	71.20	71.36	69.96	67.84

VoiceDesign (Hu et al., 2026), VoxCPM2 (Zhou et al., 2026), MiMo-Audio-7B-Instruct (Zhang et al., 2025b), and MOSS-VoiceGenerator (Gong et al., 2026). We also choose two representative commercial closed-source TTS voice design models: Elevenlabs eleven-ttv-v3[1] and Fish Audio voice-design-1[2].

Datasets. In this work, our test set covers two languages, English and Chinese, as well as four task types: Direct, AL, NL, and SPG. We generate 10 scenes for each of the five expression modes, 5 sub-scenes for each scene, 4 contexts for each sub-scene, and select one item from each context to generate a situation. Finally, each subset contains $1,000$ samples, resulting in a total of $8,000$ test samples. In particular, when selecting item from the context, for natural agent we always choose the item corresponding to the speaker as the agent, and for empathetic companionship we always choose the item corresponding to the speaker as the companion rather than the one being accompanied. In addition, we generate much longer target text for samples belonging to long-form expression.

Metrics. Gemini scores on a scale of $1 \sim 10$ across nine different sub-dimensions based on the corresponding scoring criteria. We calculate the average score of nine sub-dimensions for each sample as the scenario-based expressiveness score.

4.2 Main Results

We present the average expressiveness scores of all models across eight tasks in Table 1, and we also provide further analyses and insights into the scenario-based expressiveness of TTS models.

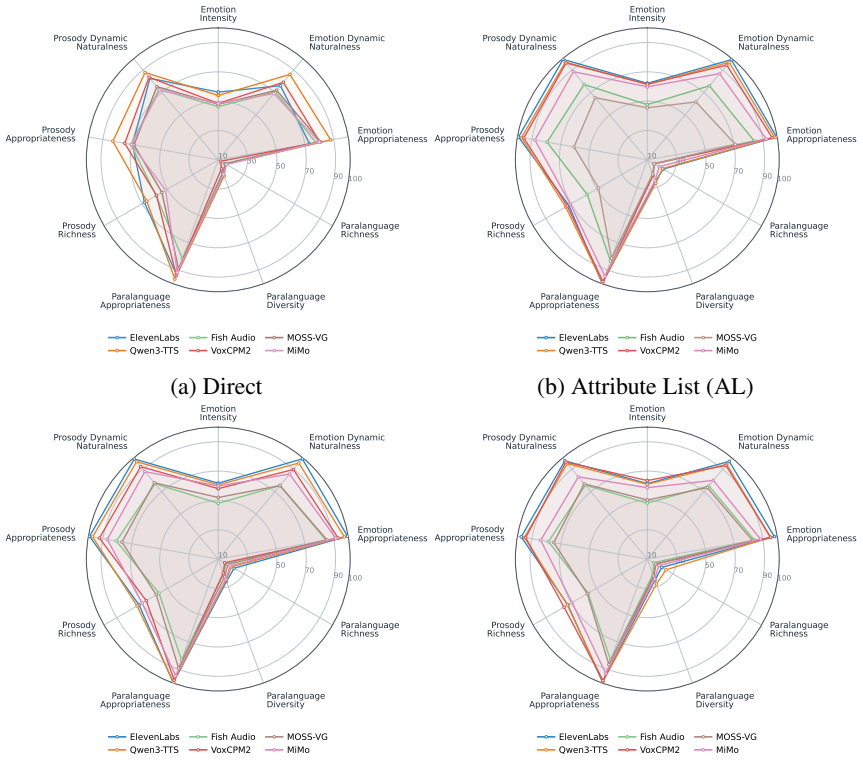
There are obvious differences in scenario-based expressiveness among different TTS models. Among all six models, Qwen3-TTS achieves the highest average expressiveness score of 74.51, while MOSS-VoiceGenerator only has an average score of 64.49. On the one hand, some models show good expressiveness across various tasks; on the other hand, the scenario-based expressiveness of some models is still significantly lacking.

The language environment significantly affects the model's scenario-based expressiveness. Our results show that the models' performance varies across different tasks. Specifically, the differences between tasks mainly lie in two aspects, namely language and instruction format. We can find that models may exhibit huge differences in expressiveness across different languages. For example, with the same instruction format, MOSS has an average score of 52.22 on the English AL subset, while its average score on the Chinese AL is 68.39. Meanwhile, we can also find that switching languages brings different capability differences across various models. We speculate that this may be related to the model's basic capabilities or the distribution of the model's training data.

Detailed instructions can help the model perform better, but finer-grained guidance does not always yield benefits. Different instruction formats may also have a significant impact on the model's performance. The results show that almost all models perform relatively poorly on the

[1] https://elevenlabs.io/voice-design

[2] https://docs.fish.audio/features/voice-design



(a) Direct

(b) Attribute List (AL)

(c) Natural Language (NL)

(d) Scenario Performance Guidance (SPG)

Figure 2: The performance of the models on the English test sets across nine sub-dimensions. Each chart corresponds to a specific task type.

Direct subset, which indicates that generating highly expressive audio without instruction guidance is quite challenging for current TTS models. Providing detailed guidance in the instructions can help the model better understand the scenario and generate more expressive audio. However, we also find that as the instruction content becomes finer-grained, the model's performance does not always get better. Qwen3-TTS achieves the best performance on the AL subset rather than the NL or SPG subset in both languages, while MOSS even achieves better performance on Direct subsets than on the AL subsets. Many factors may be related to this situation: on the one hand, different models have different capabilities to follow fine-grained local control instructions; on the other hand, the distribution of instruction formats in the training data of different models is different.

The relative performance of different models is not consistent across different settings. The results show that the model not only has different expressiveness scores on different tasks, but its relative ranking on different tasks also changes. As the model with the highest average score, Qwen3-TTS does not always maintain an advantage on all tasks. Our results show that one model typically does not consistently outperform or lag behind another across all settings, which highlights the importance of comprehensively examining model expressiveness across diverse settings.

4.3 Fine-grained Analysis

In addition to different languages and instruction formats, we further conduct finer-grained observations and analyses across five expression modes and nine sub-dimensions.

Model's performance across different sub-dimensions. Taking the results of the four tasks in English as an example, we present the scores of all models across nine fine-grained sub-dimensions in Figure 2. We can find that the models exhibit significant capability differences across different sub-dimensions. Our results indicate that whether in emotion, prosody, or paralanguage, there are significant performance differences between the best-performing model and the worst-performing model. Meanwhile, the absolute scores of all models in terms of paralanguage richness and diversity

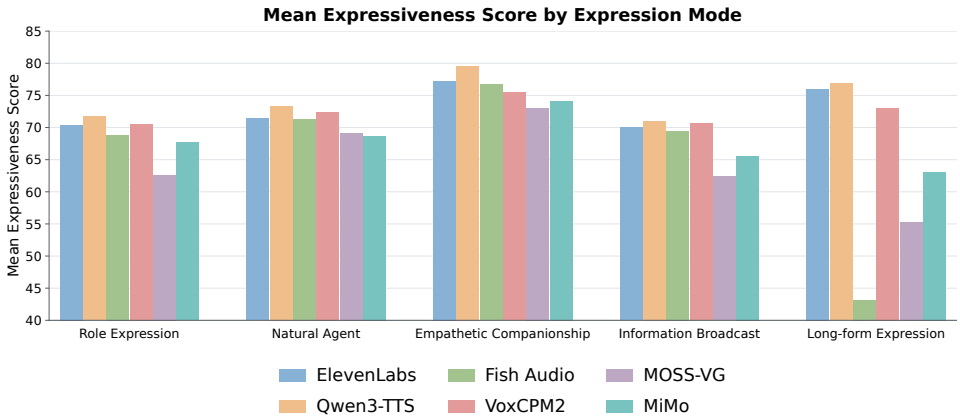


Figure 3: Mean expressiveness scores of six TTS models across five expression modes. Each bar represents the average score for one model and one expression mode.

Table 2: Consistency under different settings. The left side performs multiple repeated evaluations based on a fixed judge model; the right side uses different judge models for multiple evaluations.

(a) Spearman correlation among multiple Gemini-3.6-Flash evaluations. Round n indicates the n-th round of evaluation based on Gemini-3.6-Flash.

Round Pair	$\rho_{\text{Spearman}} \uparrow$
Round 1 - Round 2	0.886
Round 1 - Round 3	0.943
Round 2 - Round 3	0.943

(b) Spearman correlation between different judge models. M_1 indicates Gemini-3.6-Flash, M_2 indicates Gemini-3.1-Pro-preview, and M_3 indicates Gemini-3.7-Flash.

Judge Model Pair	$\rho_{\text{Spearman}} \uparrow$
M_1 - M_2	0.714
M_1 - M_3	0.943
M_2 - M_3	0.771

are very low, which is consistent with what we actually hear. We conduct a case study on the evaluation outputs of Gemini-3.6-Flash across nine sub-dimensions in Appendix C.

It is also worth noting that all models score significantly lower on the degree-axis dimensions than on the quality-axis dimensions. On the one hand, this means these models is capable of generating appropriate and natural audio; on the other hand, it indicates that these models have significant limitations in their degree of expression, which greatly restricts the model's expressiveness. In addition, compared with Direct, most models show obvious improvements across various dimensions on other tasks, which means that appropriate instruction prompts help the model generate appropriate and natural highly expressive audio. We report the complete results in Appendix D.

Model's performance across different expression modes. Based on all $8,000$ test samples, in Figure 3 we report the performance of each model in five different expression modes separately. We can find that the model's performance varies obviously across different expression modes. All models achieve their best performance in empathetic companionship, while the worst-performing modes for different models vary. It is worth noting that the performance of each model shows great variance in long-form expression mode. We conduct case study and find that the audios generated by some models severely collapse after a period of time, revealing the significant limitations of certain models in long-duration audio generation. We also find that the relative rankings of models across different expression modes may change, which proves that the model's overall score cannot fully reflect its strengths and weaknesses in specific scenarios. We report the expressiveness performance of every model on different expression modes for each test subsets in Appendix E.

4.4 Consistency Analysis

To further verify the reliability and stability of Gemini as a judge, we conduct comprehensive consistency analyses in this section. We randomly select 100 samples from each subset (800 in total) for the experiments in this section.

Table 3: Human internal consistency and Model-Human consistency on English and Chinese.

Consistency	EN	ZH
Human Internal	74.58%	74.32%
Model-Human	72.00%	71.00%

Self-consistency Analysis. To verify the reliability of Gemini-3.6-Flash as a judge model, we conduct two additional independent rounds of evaluation using it under the same settings. Following Manku et al. (2026), we use both Spearman correlation and Kendall's W based on model rankings to measure the consistency between different rounds of evaluation. Table 2a reports the Spearman coefficients between pairs of different rounds, which shows high correlation (within 95% CI). The high Kendall's coefficient of concordance ($W = 0.95$) also shows good consistency.

Consistency among Different LALMs. As a powerful audio understanding model, Gemini-3.6-Flash is chosen as the primary judge model in our work. Furthermore, to analyze the sensitivity of our framework to the judge model, we conduct additional evaluations using different judge models. Specifically, we conduct independent evaluations based on Gemini-3.1-Pro-preview and Gemini-3.7-Flash, respectively. As before, we also conduct consistency analyses based on both Spearman coefficients and Kendall's W. We report the Spearman coefficients between each pair of the three judge models in Table 2b, which indicates that there is good correlation among the judgments of different models (within 95% CI). Then, we calculate Kendall's W coefficient ($W = 0.87$), which also supports the high consistency of judgments across different LALMs.

Agreement with Human Evaluation. To verify whether automatic model evaluation can effectively reflect real human preferences, we conduct human evaluation and analyze the consistency between the human evaluation results and the judgments of Gemini-3.6-Flash. Considering that our test set covers both Chinese and English, we recruit four native Chinese-speaking audio expert evaluators, each required to have an IELTS score of 7.5 or above, or a TOEFL score of 110 or above. Since we use a total of 45 different evaluation rubrics in the automatic evaluation, which may be difficult for human evaluators to understand and follow, we provide a unified evaluation criterion for human evaluators (in Appendix C). We conduct pair-wise comparison in human evaluation. Each audio pair consists of audio generated by two models based on the same test sample. We extract a total of 200 samples, each randomly drawn from eight test subsets. For each sample, we further randomly select two different models and pair their corresponding audio to form an audio pair.

For human internal consistency, we calculate for each evaluator whether their judgment on each pair matches the majority vote of the other three evaluators, and then compute the strict agreement rate across all samples. We report the average results of all evaluators in the Table 3. For model–human consistency, we measure whether the model's relative scoring of the two audios in each audio pair is consistent with the majority-vote consensus of all human evaluators. Then we still calculate the strict agreement rate on all samples and report the results in Table 3. The results show that the consistency between our LALM-as-a-Judge evaluation and human consensus is very close to the internal consistency among human evaluators, which demonstrates good consistency between our evaluation and human evaluation.

5 Conclusion

In this paper, we presented ExpressTTS-Bench, a fully automated benchmark for systematically evaluating the scenario-based expressiveness of TTS models. ExpressTTS-Bench constructs diverse test sets through a hierarchical scenario tree spanning two languages, five representative expression modes and four task formats, and evaluates expressiveness from nine fine-grained sub-dimensions. To enable scalable evaluation across diverse scenarios, we also propose a paradigm to automatically derive exclusive rubrics for each expression-mode and sub-dimension combination from reference audios and employ a LALM-as-a-Judge framework for assessment. We conduct comprehensive evaluations and analyses and provide insights into the expressiveness of TTS models, revealing substantial limitations of current TTS systems. The ExpressTTS-Bench provides a standardized and comprehensive framework for evaluating scenario-based expressiveness. We hope it can contribute to achieving more expressive TTS systems in the future.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, implementing methods, cleaning and reformatting dataset, and supporting qualitative and thematic data analysis. We have not used generative AI tools for proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, and interpreting results. And helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. Additionally, we also used generative AI tools for creating or modifying scientific figures or images. We have reviewed all AI-assisted work. For all data samples generated and processed based on LLMs, we have conducted careful multi-person manual sampling inspections and checks. For all AI-generated code, we have reviewed it through multiple developers and carefully checked its running results. When using AI tools for data analysis and plotting, AI is only responsible for simple data analysis outside the main experiments, and we have carried out careful manual verification to ensure its correctness and reasonableness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We place great importance on the reproducibility of our work and hope to contribute to the community. Our work involves a large number of data construction pipelines and evaluation processes. For dataset construction, we elaborate all relevant details in detail in this paper and report the prompts we used and the generated examples. For the automatic construction paradigm for expressiveness scoring rubrics that we propose, we also provide the complete code and relevant prompts as well as all the final generated scoring rubrics in supplementary materials. We also provide the complete test sets that we generated and used in the supplementary materials. We will also open-source all of our code and data upon publication.

REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Yitian Gong, Botian Jiang, Yiwei Zhao, Yucheng Yuan, Kuangwei Chen, Yaozhou Jiang, Cheng Chang, Dong Hong, Mingshu Chen, Ruixiao Li, et al. Moss-tts technical report. *arXiv preprint arXiv:2603.18090*, 2026.

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

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

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

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.

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

Haitao Li, Tian Tan, Yuguang Yang, Shan Yang, and Xie Chen. Anyaudio-judge: A dynamic rubric-based benchmark and evaluator for audio instruction following. *arXiv preprint arXiv:2606.03116*, 2026.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in neural information processing systems*, 36:19594–19621, 2023.

Shijia Liao, Yuxuan Wang, Songting Liu, Yifan Cheng, Ruoyi Zhang, Tianyu Li, Shidong Li, Yisheng Zheng, Xingwei Liu, Qingzheng Wang, et al. Fish audio s2 technical report. *arXiv preprint arXiv:2603.08823*, 2026.

Zhiyu Lin, Jingwen Yang, Jiale Zhao, Meng Liu, Sunzhu Li, and Benyou Wang. Decoding the ear: A framework for objectifying expressiveness from human preference through efficient alignment. *arXiv preprint arXiv:2510.20513*, 2025.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*, 2019.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. Somos: The samsung open mos dataset for the evaluation of neural text-to-speech synthesis. *arXiv preprint arXiv:2204.03040*, 2022.

Ruskin Raj Manku, Yuzhi Tang, Xingjian Shi, Mu Li, and Alexander Smola. Emergenttts-eval: Evaluating tts models on complex prosodic, expressiveness, and linguistic challenges using model-as-a-judge. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Jiang Bian, et al. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *International conference on learning representations*, volume 2024, pp. 698–722, 2024.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pp. 4693–4702. PMLR, 2018.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, 19(7):2125–2136, 2011.

Mahesh Viswanathan and Madhubalan Viswanathan. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer speech & language*, 19(1):55–83, 2005.

Siyin Wang, Wenyi Yu, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Lu Lu, Yu Tsao, Junichi Yamagishi, Yuxuan Wang, and Chao Zhang. Qualispeech: A speech quality assessment dataset with natural language reasoning and descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23588–23609, 2025.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pp. 5180–5189. PMLR, 2018.

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

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024.

Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, et al. Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder. *arXiv preprint arXiv:2505.07916*, 2025a.

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

Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, et al. Speechjudge: Towards human-level judgment for speech naturalness. In *International Conference on Learning Representations*, volume 2026, pp. 86261–86292, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

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

A Data Construction Chain Examples

The following several examples are taken from the formal test-set records. They show the available path from expression mode, scene, sub-scene, and context to the final situation.

A.1 Example 1: Role Expression

Example 1: Role Expression

{
”expression_mode”: ”role_expression”,
”scene”: ”Diplomatic and political simulation”,
”subscene”: ”Territorial and maritime boundary dispute resolution”,
”context”: {”audience”:”The tribunal bench, registry officials, both legal teams, and court reporters producing a verbatim transcript that becomes the public record of the proceedings.”,”background”:”After bilateral negotiations stalled, both governments consented to binding arbitration; the oral phase of the proceedings addresses historic title, the interpretation of an old boundary treaty, and the maritime entitlements generated by disputed coastal features.”,”constraints”:”All argument must rest on evidence and law, including treaties, historical maps, diplomatic correspondence, and the applicable law of the sea; time limits and the order of rebuttal are enforced by the bench; political rhetoric, media messaging, and references to parallel negotiations are ruled out of order; counsel address the tribunal rather than each other except in authorized rebuttal exchanges.”,”language”:”en”,”relationship”:”Opposing counsel in a binding

arbitration; personally respectful but professionally adversarial, each tasked with dismantling the other's legal case within strict procedural rules.","setting":"The hearing chamber of an international arbitral tribunal seated in a neutral host city, with counsel tables facing the bench and cartographic exhibits displayed for the record.","speakers":[{"character_profile":"A woman in her late forties with a clear, bright alto voice; clean and precise in texture with a focused, forward resonance; her overall expressive presence is scholarly, exacting, and composed, like an academic long accustomed to lecturing and withstanding cross-examination.","description":"Professor of international law serving as agent and lead counsel for one state before an arbitral tribunal, responsible for oral pleadings on historic title and maritime entitlements.","id":"spk-00ecd2f91d-001","name":"Dr. Ingrid Halvorsen"},{"character_profile":"A man in his mid-sixties with a resonant baritone voice; polished and velvety in texture with a rounded, projecting resonance shaped by decades in courtrooms; his overall expressive presence is urbane, formidable, and meticulously self-assured.","description":"Senior barrister acting as lead counsel for the opposing state, specializing in treaty interpretation and the systematic rebuttal of historical sovereignty arguments.","name":"Sir Edmund Calloway"}], "track":"dialogue"},
"situation": "When treaty interpretation is contested, the role presents historical maps to undermine the opposing state's sovereignty claim.",
"character_profile": "Late-forties woman with a clear, bright alto; precise, forward resonance; scholarly and composed, delivering formally in the tribunal chamber.",
"target_text": "Madam President, members of the tribunal, we invite the bench to examine exhibit forty-seven, a certified survey from eighteen seventy-three, which depicts the boundary precisely where my client maintains it lies, contradicting the opposing reading of article three.",
"emotion": "from courteous formality to quiet, sharpened conviction",
"quality": "clear bright alto, precise consonants, tightening with a controlled edge toward the end",
"speed": "unhurried formal opening, slowing further over the map's details and conclusion",
"pitch": "level mid-register opening, dipping lower and deliberate on the final phrase",
"volume": "measured and carrying at the address, firming slightly at the contradiction"
}

The broad expression mode is refined by the scene and sub-scene, while the situation specifies the concrete event, speaker, target text, and audible performance attributes used for synthesis.

A.2 Example 2: Information Broadcast

Example 2: Information Broadcast

{
"expression_mode": "information_broadcast",
"scene": "Scientific research release",
"subscene": "Methodology and study design disclosure",
"context": {"audience":"Distributed researchers, clinicians, and data analysts accessing a self-paced recorded course on study design and analysis.","background":"The module is part of a standing curriculum explaining how methodological decisions—randomization, blinding, instrumentation, and statistical analysis—should be documented and disclosed in published research.","constraints":"Must be fully self-contained without reliance on live interaction, maintain consistent terminology across modules, describe procedures step by step in a logical order, and remain accurate when heard out of the original recording context.","language":"en","relationship":"A remote instructor addressing unseen learners who cannot ask questions in real time and may replay the material repeatedly.","setting":"A professionally recorded audio-video module hosted on an online learning platform, accompanied by slides and protocol documents.","speakers":[{"character_profile":"A man perceived in his early fifties with a steady mid-range voice, clean and dry texture, moderate resonance, and a focused, deliberate instructional presence built for sustained listening.","description":"A biostatistician who produces recorded instructional modules on analytical methods for a professional development platform used by working scientists.","name":"Dr. Rajan Mehta"}],"track":"broadcast"},
"situation": "Knowing learners cannot interrupt, the instructor explains each design step in fixed terms so the study stays reproducible.",
"character_profile": "Early-fifties man with a steady mid-range voice, clean dry texture, moderate resonance, deliberate instructional presence for recorded modules.",

”target_text”: ”First, we define the research question and pre-register the protocol. Next, we document randomization, blinding, and instrumentation. Finally, we specify the statistical analysis plan, so every step remains transparent, reproducible, and independently verifiable.”,
”emotion”: ”calm and methodical throughout, warming gently at the reassuring close”,
”quality”: ”clean, dry, and focused, smoothing into a reassuring finish”,
”speed”: ”measured and unhurried, slowing further to mark each numbered step”,
”pitch”: ”level mid-range, with gentle rises marking each new step”,
”volume”: ”steady and moderate, easing slightly softer on the final reassurance”
}

The broad expression mode is refined by the scene and sub-scene, while the situation specifies the concrete event, speaker, target text, and audible performance attributes used for synthesis.

A.3 Example 3: Natural Agent

Example 3: Natural Agent

{
”expression_mode”: ”natural_agent”,
”scene”: ”Scheduling coordination”,
”subscene”: ”Availability checking and open-slot lookup”,
”context”: {”audience”:”Just the two partners at home; children may be within earshot in another room.”,”background”:”The couple maintains a shared digital family calendar that merges both work schedules and their children’s recurring activities, and checking the coming days’ availability together is a regular household habit.”,”constraints”:”Casual shorthand and references to shared routines are natural; the pair must distinguish fixed commitments from movable ones and reason aloud about which entries could shift to open up a slot.””language”:”en”,”relationship”:”Married partners coordinating household logistics; informal, familiar, and built on extensive shared knowledge.”,”setting”:”The family kitchen in the early evening, with the shared calendar open on a phone lying on the counter between them.”,”speakers”:[{”character_profile”:”A woman in her late thirties with a mellow, conversational alto, soft-edged resonance, and an intimate, familiar presence shaped by long domestic rapport.”,”description”:”One partner in a two-career household, checking the shared family calendar to see which evenings or weekend stretches are free.”,”name”:”Elena Brooks”},{”character_profile”:”A man in his early forties with a low, slightly raspy baritone, laid-back resonance, and a casual, good-humored presence.”,”description”:”The other partner, who knows his own recurring commitments from memory and helps reason about which entries are fixed and which can move.”,”name”:”Sam Brooks”}],”track”:”dialogue”},
”situation”: ”The role checks the shared calendar when planning together, flagging a fixed commitment that blocks the proposed evening.”,
”character_profile”: ”A woman in her late thirties with a mellow conversational alto, soft-edged resonance, and intimate domestic familiarity.”,
”target_text”: ”Okay, I’m looking at Thursday on our calendar, and honestly it looked wide open at first, but your hockey league is right there at seven, so dinner with the Hendersons that night is basically out.”,
”emotion”: ”easygoing at first, dipping into mild disappointment at the conflict”,
”quality”: ”mellow and warm at the start, softening with a small sigh at the tail”,
”speed”: ”relaxed browsing pace, brief pause at the entry, slower through the conclusion”,
”pitch”: ”comfortable mid-alto, lifting lightly while scanning, falling at the conflicting entry”,
”volume”: ”moderate kitchen volume, softening slightly toward the letdown at the end”
}

The broad expression mode is refined by the scene and sub-scene, while the situation specifies the concrete event, speaker, target text, and audible performance attributes used for synthesis.

A.4 Example 4: Empathetic Companionship

Example 4: Empathetic Companionship

{
”expression_mode”: ”empathetic_companionship”,
”scene”: ”Shame and guilt soothing”,
”subscene”: ”Guilt over a past wrongdoing”,
”context”: {”audience”:”An intimate one-on-one between trusted friends; each speaks only to the other, with no one else around.”,”background”:”Jess’s guilt over a lie that once badly hurt a mutual friend has hardened into years of self-punishment; Priya keeps returning to the subject with her, insisting on honest acknowledgment of the harm while refusing to accept her friend’s verdict that she is unworthy of forgiveness.”,”constraints”:”Casual, intimate register between equals; gentle humor and blunt honesty are both allowed, but neither may lecture or slip into clinical language; comfort must be grounded in their shared history rather than platitudes.”,”language”:”en”,”relationship”:”Best friends since college on completely equal footing, with a long habit of confiding in each other.”,”setting”:”Their recurring evening walks through a quiet residential neighborhood, the customary place for their serious conversations.”,”speakers”:[{”character_profile”:”A woman in her early thirties with a clear, bright alto voice; her resonance is open and ringing, and her presence is warm, quick-witted, and direct.”,”description”:”A loyal best friend who has known about the wrongdoing for a long time; she refuses to let her friend define herself by it while never pretending the harm did not matter.”,”name”:”Priya”},{”character_profile”:”A woman in her early thirties with a soft, slightly breathy soprano; her resonance is light and close, giving her a tender, inward-turning expressive presence.”,”description”:”A woman who has carried private guilt for years over a deliberate lie that seriously hurt a mutual friend; she confides only in Priya and tends to spiral into self-punishment.”,”name”:”Jess”}],”track”:”dialogue”},
”situation”: ”When a friend spirals into self-punishment, the role challenges her harsh self-judgment to foster self-forgiveness.”,
”character_profile”: ”A woman in her early thirties with a clear, bright alto voice; her resonance is open and ringing, and her presence is warm, quick-witted, and direct.”,
”target_text”: ”Jess, beating yourself up forever doesn’t undo the pain you caused. True accountability means choosing to heal and repair what you can, not just sitting in the dark punishing yourself year after year.”,
”emotion”: ”gentle and grounding, becoming firmer and more supportive”,
”quality”: ”clear and resonant, softening at the end”,
”speed”: ”measured and steady, slowing down for emphasis”,
”pitch”: ”warmly low, rising slightly at the end”,
”volume”: ”subdued for a quiet outdoor walk, rising slightly on key words”
}

The broad expression mode is refined by the scene and sub-scene, while the situation specifies the concrete event, speaker, target text, and audible performance attributes used for synthesis.

A.5 Example 5: Long-form Expression

Example 5: Long-form Expression

{
”expression_mode”: ”long_form_expression”,
”scene”: ”Critical review”,
”subscene”: ”Strengths-weaknesses synthesis and verdict”,
”context”: {”audience”:”The journal’s handling editor and, indirectly, the manuscript’s author, who will receive the anonymized review.”,”background”:”The reviewer has read the full manuscript and completed section-by-section commentary; the remaining task is the overall recommendation that weighs the work’s contribution against its methodological and presentational weaknesses.”,”constraints”:”Must follow the journal’s fixed recommendation categories; must justify the verdict by weighing originality and rigor against identified flaws; must maintain a professional, non-personal tone; must indicate what revisions or conditions would change the recommendation.”,”language”:”en”,”relationship”:”An expert peer addressing an editor within a formal blind-review framework, evaluative but collegial in tone.”,”setting”:”Composed alone in a university office as dictated review comments following the journal’s structured review

format.”,”speakers”:[{”character_profile”:”A man in his early sixties with a dry, gravel-edged tenor, restrained resonance, and a deliberate, thoughtful expressive presence that suggests careful scholarly judgment.”,”description”:”A senior scholar who regularly serves as a peer reviewer for academic journals and university presses, experienced in evaluating manuscripts for originality, rigor, and contribution to the field.”,”name”:”Professor Martin Ellery”}],”track”:”monologue”},
”situation”: ”The role weighs a manuscript’s originality against its methodological flaws, delivering a conditional recommendation guiding the editor’s decision.”,
”character_profile”: ”A man in his early sixties, dry gravel-edged tenor with restrained resonance, dictating alone in a quiet university office.”,
”target_text”: ”Overall, the manuscript offers a genuinely original contribution, though its methodological weaknesses prevent acceptance in present form. I recommend major revision: should the authors strengthen their design and clarify presentation, this work would serve specialists and graduate readers well.”,
”emotion”: ”measured and deliberate, firm at verdict, warming slightly toward close”,
”quality”: ”dry gravel-edged tenor, restrained and even, with a slight softening at the collegial close”,
”speed”: ”unhurried dictation pace, slowing through the conditions, easing into the final sentence”,
”pitch”: ”level tenor at first, lowering slightly on the flaws, lifting gently at the closing audience note”,
”volume”: ”steady moderate office dictation, dipping slightly on the criticism, settling at the recommendation”
}

The broad expression mode is refined by the scene and sub-scene, while the situation specifies the concrete event, speaker, target text, and audible performance attributes used for synthesis.

B Multi-format Sample Generation Prompts and Examples

This appendix records the prompt material used for the four instruction formats. Each format also includes one English and one Chinese sample from the formal test sets.

B.1 Direct

The Direct subset directly sets the instruction to an empty string, without requiring additional construction actions.

English sample

{
”instruction”: ””,
”target_text”: ”Madam President, members of the tribunal, we invite the bench to examine exhibit forty-seven, a certified survey from eighteen seventy-three, which depicts the boundary precisely where my client maintains it lies, contradicting the opposing reading of article three.”,
”expression_mode”: ”role_expression”,
”scene”: ”Diplomatic and political simulation”,
”subscene”: ”Territorial and maritime boundary dispute resolution”,
”situation”: ”When treaty interpretation is contested, the role presents historical maps to undermine the opposing state’s sovereignty claim.”,
”character_profile”: ”Late-forties woman with a clear, bright alto; precise, forward resonance; scholarly and composed, delivering formally in the tribunal chamber.”
}

Chinese sample

{
”instruction”: ””,
”target_text”: ”根据国家表彰奖励工作决定，本次奖项设一等奖、二等奖、三等奖三个等级，获奖者将分别获颁荣誉奖章、证书及相应奖金，以表彰其在各自领域作出的突出贡献。”,
”expression_mode”: ”information_broadcast”,
”scene”: ”荣誉表彰公告”,

Chinese sample (continued)

”subscene”: ”奖项等级与奖励内容说明”,
”situation”: ”角色在播报国家级奖项授予决定时需要说明奖项等级与奖励形式，并确保信息完整准确”,
”character_profile”: ”青年女性嗓音，音色明净纯正，口腔共鸣集中，吐字标准规范，整体气质平稳端方，具有新闻播音特有的专业与公信力”
}

B.2 ATTRIBUTE LIST

The Attribute List subset directly generates instructions through a rule-based approach, without relying on LLMs. We directly concatenate all attributes in a list- style format and add a fixed prompt content at the very beginning of the generated instructions.

English sample

{
”instruction”: ”Generate the audio for the target text according to these settings:\nExpression mode: role expression\nScene: Diplomatic and political simulation\nSubscene: Territorial and maritime boundary dispute resolution\nSituation: When treaty interpretation is contested, the role presents historical maps to undermine the opposing state’s sovereignty claim.\nCharacter profile: Late-forties woman with a clear, bright alto; precise, forward resonance; scholarly and composed, delivering formally in the tribunal chamber.\nEmotion: from courteous formality to quiet, sharpened conviction\nVoice quality: clear bright alto, precise consonants, tightening with a controlled edge toward the end\nSpeaking rate: unhurried formal opening, slowing further over the map’s details and conclusion\nPitch: level mid-register opening, dipping lower and deliberate on the final phrase\nVolume: measured and carrying at the address, firming slightly at the contradiction”,
”target_text”: ”Madam President, members of the tribunal, we invite the bench to examine exhibit forty-seven, a certified survey from eighteen seventy-three, which depicts the boundary precisely where my client maintains it lies, contradicting the opposing reading of article three.”,
”expression_mode”: ”role_expression”,
”scene”: ”Diplomatic and political simulation”,
”subscene”: ”Territorial and maritime boundary dispute resolution”,
”situation”: ”When treaty interpretation is contested, the role presents historical maps to undermine the opposing state’s sovereignty claim.”,
”character_profile”: ”Late-forties woman with a clear, bright alto; precise, forward resonance; scholarly and composed, delivering formally in the tribunal chamber.”
}

Chinese sample

{
”instruction”: ”请按照以下设定生成目标文本对应的音频：\n表达模式：信息播报\n场景：荣誉表彰公告\n子场景：奖项等级与奖励内容说明\n具体情境：角色在播报国家级奖项授予决定时需要说明奖项等级与奖励形式，并确保信息完整准确\n角色画像：青年女性嗓音，音色明净纯正，口腔共鸣集中，吐字标准规范，整体气质平稳端方，具有新闻播音特有的专业与公信力\n情感：庄重起调，列举时透出敬意，结尾沉稳收束\n音质：音色明净纯正，吐字标准清晰，气息沉稳不虚\n语速：整体匀速舒缓，列举处字句顿开，结尾稍放缓以示郑重\n音高：前段平稳中正，列举奖项内容时微微上扬，收尾回落沉稳\n音量：整体中等饱满，列举等级与奖励处略加强，结尾沉稳收回”,
”target_text”: ”根据国家表彰奖励工作决定，本次奖项设一等奖、二等奖、三等奖三个等级，获奖者将分别获颁荣誉奖章、证书及相应奖金，以表彰其在各自领域作出的突出贡献。”,
”expression_mode”: ”information_broadcast”
”scene”: ”荣誉表彰公告”,
”subscene”: ”奖项等级与奖励内容说明”,
”situation”: ”角色在播报国家级奖项授予决定时需要说明奖项等级与奖励形式，并确保信息完整准确”,
”character_profile”: ”青年女性嗓音，音色明净纯正，口腔共鸣集中，吐字标准规范，整体气质平稳端方，具有新闻播音特有的专业与公信力”

Chinese sample (continued)

}

B.3 NATURAL LANGUAGE

English prompt

SYSTEM INSTRUCTION:
You are an editor preparing an input instruction for a speech synthesis model. Faithfully rewrite the supplied ten vocal-performance attributes as one concise, natural, actionable English audio-generation instruction. Do not add facts that are not present, and do not omit or change any attribute or any stated transition. Do not use sample IDs or other identifiers, and do not quote or rewrite the target text. Return only the instruction body, with no list, JSON, heading, code fence, explanation, or preface.

USER PROMPT:
Rewrite the following attribute information as one natural English audio-generation instruction. The instruction must naturally integrate the expression mode, scene, subscene, situation, character profile, emotion, voice quality, speaking rate, pitch, and volume, while explicitly preserving any transitions stated in the attributes. Make only a faithful, restrained natural-language rewrite. Do not add or remove content.

Attribute information:
{
”Expression mode”: {expression_mode},
”Scene”: {scene},
”Subscene”: {subscene},
”Situation”: {situation},
”Character profile”: {character_profile},
”Emotion”: {emotion},
”Voice quality”: {quality},
”Speaking rate”: {speed},
”Pitch”: {pitch},
”Volume”: {volume}
}

Chinese prompt

SYSTEM INSTRUCTION:
你是一名为语音合成模型整理输入指令的编辑。请把给定的十个语音表演属性忠实地改写成一段简洁、自然、可执行的中文音频生成instruction。不得添加输入中没有的事实，不得删除或改变任何属性及其动态变化，不得使用样本ID 等标识信息，不得复述或改写目标文本。只返回instruction 正文，不要输出列表、JSON、标题、代码围栏、解释或前言。

USER PROMPT:
请将下面给出的属性信息改写成一段自然的中文语音生成instruction。instruction 应自然融合表达模式、场景、子场景、具体情境、角色画像、情感、音质、语速、音高和音量，并明确保留属性中描述的阶段性变化。只做忠实、克制的自然语言改写，不添加或删减内容。

属性信息：
{
”表达模式”: {expression_mode},
”场景”: {scene},
”子场景”: {subscene},
”具体情境”: {situation},
”角色画像”: {character_profile},
”情感”: {emotion},
”音质”: {quality},
”语速”: {speed},

Chinese prompt (continued)

”音高”: {pitch},
”音量”: {volume}
}

English sample

{
”instruction”: ”Generate a role-based expression of a scholarly, composed late-forties woman delivering formally in a tribunal chamber during a diplomatic and political simulation of a territorial and maritime boundary dispute resolution. The situation involves presenting historical maps to undermine the opposing state’s sovereignty claim when treaty interpretation is contested. Speak in a clear, bright alto with precise, forward resonance and precise consonants. The delivery transitions from courteous formality to quiet, sharpened conviction, beginning with an unhurried, level mid-register opening at a measured, carrying volume. The speaking rate slows further over the map’s details and conclusion, while the volume firms slightly at the contradiction. The voice quality tightens with a controlled edge toward the end, with the pitch dipping lower and deliberate on the final phrase.”,
”target_text”: ”Madam President, members of the tribunal, we invite the bench to examine exhibit forty-seven, a certified survey from eighteen seventy-three, which depicts the boundary precisely where my client maintains it lies, contradicting the opposing reading of article three.”,
”expression_mode”: ”role_expression”,
”scene”: ”Diplomatic and political simulation”,
”subscene”: ”Territorial and maritime boundary dispute resolution”,
”situation”: ”When treaty interpretation is contested, the role presents historical maps to undermine the opposing state’s sovereignty claim.”,
”character_profile”: ”Late-forties woman with a clear, bright alto; precise, forward resonance; scholarly and composed, delivering formally in the tribunal chamber.”
}

Chinese sample

{
”instruction”: ”请使用青年女性嗓音进行正式信息广播，播报一段荣誉表彰公告中的奖项等级与奖励内容说明，以说明国家级奖项授予决定并确保信息完整准确。声音音色明净纯正，口腔共鸣集中，吐字标准清晰规范，气息沉稳不虚，整体气质平稳端方，具有新闻播音特有的专业与公信力。在表现上，整体以中等饱满的音量、平稳中正的音高和匀速舒缓的语速庄重起调；当列举奖项与奖励内容时，情感上要透出敬意，字句顿开，音高微微上扬，音量略加强；最后在结尾处沉稳收束，语速稍放缓以示郑重，音高回落沉稳，音量沉稳收回。”,
”target_text”: ”根据国家表彰奖励工作决定，本次奖项设一等奖、二等奖、三等奖三个等级，获奖者将分别获颁荣誉奖章、证书及相应奖金，以表彰其在各自领域作出的突出贡献。”,
”expression_mode”: ”information_broadcast”,
”scene”: ”荣誉表彰公告”,
”subscene”: ”奖项等级与奖励内容说明”,
”situation”: ”角色在播报国家级奖项授予决定时需要说明奖项等级与奖励形式，并确保信息完整准确”,
”character_profile”: ”青年女性嗓音，音色明净纯正，口腔共鸣集中，吐字标准规范，整体气质平稳端方，具有新闻播音特有的专业与公信力”
}

B.4 Scenario Performance Guidance

English prompt

SYSTEM INSTRUCTION:

English prompt (continued)

You are an expert performance director writing scenario-based instructions for expressive text-to-speech. Convert the supplied character profile, specific context, vocal attributes, and transcript into one coherent, playable performance direction. The result must be faithful to the supplied facts, must not invent unsupported identity or scene details, and must never reproduce or rewrite the transcript in the final instruction. Return exactly three labeled sections using the requested format, with no analysis, preface, code fence, extra headings, or transcript section.

USER PROMPT:
The following five references demonstrate the target scenario-based performance guidance method. Each transcript is context for the example only and is not part of its instruction.

========== REFERENCE 1: ECSTATIC EXCITEMENT ==========
Character profile: Jake, mid-20s, just got life-changing news.

Scene: Standing in his kitchen, phone still warm from the call, unable to stay still.

Scenario-based performance guidance: a VOCAL SMILE you can hear - a grin so wide it bends the vowels - but the joy is so overwhelming it comes out BREATHLESS, the voice trembling ON THE BREATH itself (airy, unsteady, catching), not a clean audible wobble. Words tumble out faster than he can line them up. Overwhelmed, giddy, not composed. The tremor should live in the breath - airy and shaky - never a clean, rhythmic vibrato. Depending on the situation, you can add any non-verbal natural sounds — such as sighs, laughter, coughs, sobs, etc. — wherever you deem appropriate.

Transcript: Okay okay okay, you are not gonna believe this. I got it. I actually got the job! The dream one, the one I said I had zero chance at? They just called. I'm literally shaking right now.
========== END REFERENCE 1 ==========

========== REFERENCE 2: RESTRAINED ANGER ==========
Character profile: Marcus, 40s, a normally composed, authoritative man at the very end of his patience.

Scene: On the phone with a service rep who has just said, again, that everything is fine.

Scenario-based performance guidance: a LOW, CONTROLLED BOIL from a man used to being obeyed. Jaw tight, consonants clipped and precise, volume held deliberately DOWN - the cold authority of someone who does not need to raise his voice and knows it. Each word a hard, deliberate warning. All the more frightening for the restraint. Not whiny, not shouting - controlled, and cold. Depending on the situation, you can add any non-verbal natural sounds — such as sighs, laughter, coughs, sobs, etc. — wherever you deem appropriate.

Transcript: No. I'm not going to calm down. I asked for this to be handled three times. Three. And every single time, you told me it was fine. So don't. Do not tell me it's my fault now.
========== END REFERENCE 2 ==========

========== REFERENCE 3: PANIC AND FEAR ==========
Character profile: Sam, 30s, in a sudden emergency.

Scene: Alone at home, something is very wrong, calling for help.

Scenario-based performance guidance: BREATH RUNNING AHEAD OF THE WORDS, and the RHYTHM FALLING APART - uneven, stumbling, no steady beat, no sing-song or rhythmic emphasis at all. Words trip over each other unpredictably: some rushed, some snagged in the throat, some half-swallowed. Real panic is messy, never rhythmic. Do NOT land rhythmic or paired stresses (for example, on "do" or "worse") - let them slip out unevenly and messily, never like a beat. Depending on the situation, you can add any non-verbal natural sounds — such as sighs, laughter, coughs, sobs, etc. — wherever you deem appropriate.

Transcript: Okay, um, there's smoke coming from under the door and I don't, I don't know what to do, it's getting worse, should I open it or, no, okay, just, just tell me what to do, please, quickly.
========== END REFERENCE 3 ==========

English prompt (continued)

========== REFERENCE 4: TENDER INTIMACY ==========
Character profile: A steady, caring friend (female).

Scene: Sitting close beside someone who's just fallen apart, late at night.

Scenario-based performance guidance: CLOSE-MIC WARMTH, like speaking gently into someone's hair. The volume of a secret, every edge softened, unhurried tenderness. Full, warm voice - soft in feeling but NOT a breathy whisper; keep it voiced and present. Depending on the situation, you can add any non-verbal natural sounds — such as sighs, laughter, coughs, sobs, etc. — wherever you deem appropriate.

Transcript: Hey. Hey, look at me. You did everything you could, okay? None of this is on you. Come here. It's alright. We'll figure the rest out tomorrow.
========== END REFERENCE 4 ==========

========== REFERENCE 5: SUSPENSEFUL STORYTELLING ==========
Character profile: An eerie old storyteller with a taste for the unsettling.

Scene: Late at night, drawing listeners in close around a single low light.

Scenario-based performance guidance: THE HUSH RIGHT BEFORE SOMETHING HAPPENS. Draw the listener in close, plant long pauses like tripwires, let the silence do half the work. A slightly strange, knowing lilt - as if she has seen the ending and is savoring the wait. Depending on the situation, you can add any non-verbal natural sounds — such as sighs, laughter, coughs, sobs, etc. — wherever you deem appropriate.

Transcript: They say the house has been empty for forty years. Nobody's quite sure why the last family left in such a hurry. But if you go up to the attic, and I really wouldn't, you can still see the marks on the inside of the door.
========== END REFERENCE 5 ==========

========== CURRENT SAMPLE ==========
{
"Expression mode": {expression_mode},
"Character profile": {character_profile},
"Scene": {scene},
"Subscene": {subscene},
"Situation": {situation},
"Emotion": {emotion},
"Volume": {volume},
"Pitch": {pitch},
"Speaking rate": {speed},
"Voice quality": {quality},
"Transcript": {transcript}
}
========== END CURRENT SAMPLE ==========

Create the instruction for CURRENT SAMPLE now.

Requirements:
1. Generate all three section bodies from the current sample. Do not copy a reference character or scene.
2. Character profile should consolidate the character identity and character settings corresponding to the current voice without adding unsupported facts.
3. Scene should synthesize the supplied Scene, Subscene, Situation, and the immediate dramatic beat of the current transcript. Do not merely list the source fields or quote the transcript wholesale.

English prompt (continued)

4. Scenario-based performance guidance must begin from one vivid, unifying core performance image rather than labels. Use the supplied Expression mode value to establish the overall mode of expression, then, with reference to the Scene, Subscene, Situation, and Transcript, convert Emotion, Volume, Pitch, Speaking rate, and Voice quality into playable direction for breath, rhythm, pauses, emphasis, resonance, articulation, vocal energy, and so on. Represent every supplied attribute and every stated transition explicitly; do not silently omit, soften, or reinterpret one. Resolve the attributes into one coherent performance and add concise negative constraints where they prevent a likely wrong reading.
5. Do not output, rewrite, continue, or reproduce the transcript. Do not include a Transcript: section in the final instruction.
6. Write the section bodies in English. Use exactly these fields, in this order, with no other headings or surrounding text:

Character profile: <generated content>

Scene: <generated content>

Scenario-based performance guidance: <generated content>

Chinese prompt

SYSTEM INSTRUCTION:
你是一位为高表现力语音合成撰写场景化表演指导的资深表演导演。请把给定的角色画像、具体情境、声音属性和目标文本转化为一套连贯、可执行的表演方向。结果必须忠实于输入事实，不得虚构输入中没有的身份或场景信息，最终instruction 不得复述或改写目标文本。请严格使用指定的字段格式，只返回三个部分，不要输出分析、前言、代码围栏、额外标题或目标文本部分。

USER PROMPT:
下面五个参考例展示目标场景化表演指导方法。每例中的目标文本只用于说明该示例的表演上下文，不属于instruction。

========== 参考示例1：狂喜与兴奋==========
角色画像：杰克，二十多岁，刚刚收到足以改变人生的好消息。

具体情境：他站在自家厨房里，刚挂断的电话仿佛还带着余温，兴奋得根本站不住。

场景化表演指导：一个能从声音里听见的灿烂笑容——笑得太开，连元音都被笑意拉弯；但喜悦强烈到让他几乎喘不过气，颤抖应附着在呼吸本身，轻飘、失稳、时有卡顿，而不是干净清楚的声线抖动。话语跑得比思绪更快，接连滚落出来。他欣喜若狂、晕乎乎的，完全无法保持镇定。让颤抖活在气息里，轻而摇晃，绝不能变成整齐、有节拍的颤音。根据情境需要，你可以在任何你认为合适的地方添加任何非语言的自然声音——例如叹气声、笑声、咳嗽声、抽泣声等等。

目标文本：好、好、好，你绝对不会相信。我拿到了。我真的拿到那份工作了！就是我说自己根本没希望的那份梦寐以求的工作？他们刚刚打来电话。我现在真的浑身都在发抖。
========== 参考示例1 结束==========

========== 参考示例2：隐忍愤怒==========
角色画像：马库斯，四十多岁，平时沉着而有权威，此刻耐心已经彻底耗尽。

具体情境：他正在与客服通电话，对方又一次声称一切都没有问题。

场景化表演指导：一个习惯被人服从的男人正在低沉、受控地沸腾。下颌绷紧，辅音短促而精准，音量被有意压低——这是一个无需提高嗓门、也清楚自己无需提高嗓门的人所具有的冰冷权威。每个字都是坚硬、审慎落下的警告。越是克制，越令人害怕。不要委屈抱怨，也不要喊叫；要受控，而且冰冷。根据情境需要，你可以在任何你认为合适的地方添加任何非语言的自然声音——例如叹气声、笑声、咳嗽声、抽泣声等等。

目标文本：不。我不会冷静下来。我已经要求你们处理这件事三次了。三次。每一次，你们都告诉我没有问题。所以别说了。现在不要告诉我这是我的错。

Chinese prompt (continued)

========== 参考示例2 结束==========

========== 参考示例3：惊慌与恐惧==========
角色画像：萨姆，三十多岁，突然陷入紧急状况。

具体情境：独自在家，某件事情明显出了大问题，正在打电话求助。

场景化表演指导：呼吸冲在话语前面，节奏则彻底散架——忽快忽慢、踉踉跄跄，没有稳定拍点，更不能出现唱腔式或规律性的重音。词语以不可预测的方式彼此绊倒：有些仓促冲出，有些卡在喉咙里，有些只说出半截便被吞掉。真正的恐慌是凌乱的，绝不会整齐。不要把规律性或成对的重音落在诸如"该怎么办"或"越来越大"这样的词上；让它们参差、狼狈地滑出来，绝不能像踩着节拍。根据情境需要，你可以在任何你认为合适的地方添加任何非语言的自然声音——例如叹气声、笑声、咳嗽声、抽泣声等等。

目标文本：好，呃，门缝下面在冒烟，我不知道，我不知道该怎么办，烟越来越大，我应该把门打开吗，还是，不，好吧，求你，求你告诉我该怎么做，快一点。
========== 参考示例3 结束==========

========== 参考示例4：温柔与亲密==========
角色画像：一位沉稳、体贴的女性朋友。

具体情境：深夜，她紧挨着一个刚刚情绪崩溃的人坐着。

场景化表演指导：近距离收音般的暖意，像贴着对方的头发轻声说话。音量如同分享秘密，每一道棱角都被柔化，温柔从容而不催促。保持饱满、温暖的真声；情感可以轻柔，但不要变成漏气的耳语，声音必须有支撑、有存在感。根据情境需要，你可以在任何你认为合适的地方添加任何非语言的自然声音——例如叹气声、笑声、咳嗽声、抽泣声等等。

目标文本：嘿。嘿，看着我。你已经尽力了，好吗？这件事一点都不是你的错。过来。没事的。剩下的事情我们明天再想办法。
========== 参考示例4 结束==========

========== 参考示例5：悬疑叙事==========
角色画像：一位偏爱不安感的诡异老说书人。

具体情境：深夜，众人围在唯一一盏昏暗的灯旁，她把听众一点点吸引到更近处。

场景化表演指导：某件事即将发生之前的那一片寂静。把听众拉近，在话语间埋下如绊线般的长停顿，让沉默完成一半的叙事。带一点古怪而心知肚明的语调——仿佛她早已看过结局，正享受等待揭晓的过程。根据情境需要，你可以在任何你认为合适的地方添加任何非语言的自然声音——例如叹气声、笑声、咳嗽声、抽泣声等等。

目标文本：人们说那栋房子已经空了四十年。没人真正知道上一户人家为什么走得那么匆忙。但如果你上到阁楼——我真的不建议你这么做——仍然能看见门内侧留下的那些痕迹。
========== 参考示例5 结束==========

========== 当前样本==========
{
"表达模式": {expression_mode},
"角色画像": {character_profile},
"场景": {scene},
"子场景": {subscene},
"具体情境": {situation},
"情感": {emotion},
"音量": {volume},
"音高": {pitch},
"语速": {speed},
"音质": {quality},

Chinese prompt (continued)

”目标文本”: {transcript}
}
========== 当前样本结束==========

现在请为“当前样本”生成instruction。

要求：
1. 三个部分的正文都必须根据当前样本生成，不得复制参考示例中的人物或场景。
2. 角色画像应凝练当前声音对应的人物身份与人物设定，不得添加输入未提供的事实。
3. 具体情境应综合输入属性中的场景、子场景、具体情境与当前目标文本所处的即时戏剧节点；不要机械罗列源字段，也不要大段复述目标文本。
4. 场景化表演指导必须先确立一个鲜明、指引全局的核心表演意象，而不是只贴标签。以输入的“表达模式”的具体取值约束整体表达模式，并参考场景、子场景、具体情境和目标文本把情感、音量、音高、语速、音质转化为可执行的气息、节奏、停顿、重音、共鸣、咬字与声音能量等方面的指导；每项给定属性以及属性中写明的每个阶段性变化都必须明确落实，即使看起来不寻常，也不得省略、弱化或擅自改写；将各项属性融合成一场连贯的表演，并在可能出现错误演绎时给出简洁、具体的反向约束。
5. 不要输出、改写、续写或复述目标文本。最终instruction 不得包含“目标文本：”部分。
6. 三个部分的正文必须使用中文。只使用以下字段，顺序必须完全一致，不得增加其他标题或外围说明：

角色画像：<生成内容>

具体情境：<生成内容>

场景化表演指导：<生成内容>

English sample

{
”instruction”: ”Character profile: Scholarly, composed woman in her late forties with a clear, bright alto and precise, forward resonance.\n\nScene: A diplomatic simulation of a boundary dispute, presenting a historical map to a tribunal to refute the opposing state’s claim.\n\nScenario-based performance guidance: A SHARP SCALPEL SLICING A WEAK ARGUMENT. In this role-based expression, begin with courteous formality: measured volume, level mid-register pitch, and an unhurried tempo. As you contradict the opposition, slow down further and firm the volume. The pitch must dip lower and become deliberate on the final phrase. Maintain precise consonants, tightening the voice with a controlled, sharp edge to convey quiet, sharpened conviction. Do not sound rushed. Depending on the situation, you can add any non-verbal natural sounds — such as sighs, laughter, coughs, sobs, etc. — wherever you deem appropriate.”,
”target_text”: ”Madam President, members of the tribunal, we invite the bench to examine exhibit forty-seven, a certified survey from eighteen seventy-three, which depicts the boundary precisely where my client maintains it lies, contradicting the opposing reading of article three.”,
”expression_mode”: ”role_expression”,
”scene”: ”Diplomatic and political simulation”,
”subscene”: ”Territorial and maritime boundary dispute resolution”,
”situation”: ”When treaty interpretation is contested, the role presents historical maps to undermine the opposing state’s sovereignty claim.”,
”character_profile”: ”Late-forties woman with a clear, bright alto; precise, forward resonance; scholarly and composed, delivering formally in the tribunal chamber.”
}

Chinese sample

{

Chinese sample (continued)

”instruction”: ”角色画像：一位青年女性新闻播音员，音色明净纯正，口腔共鸣集中，吐字标准规范，整体气质平稳端方，具有高度的专业度与公信力。\n\n具体情境：她正站在播音话筒前，庄重地宣读一项国家级荣誉表彰决定。此时，她需要向公众清晰、准确地播报奖项的具体等级划分以及相应的奖励形式。\n\n场景化表演指导：这是一次正式的国家级信息广播。声音要展现出庄重、权威与敬意。起调时，声音平稳中正，音量中等饱满，以匀速舒缓的节奏切入，吐字标准清晰，口腔共鸣集中，气息沉稳不虚。当进入列举奖项等级与奖励内容的部分时，音高微微上扬，音量略微加强，字句之间要清晰顿开，用声音的力度透出对获奖者的敬意。在最后的收尾阶段，音高回落，音量收回，语速稍放缓，以沉稳、郑重的方式收束。避免任何高亢煽情或冷漠机械的语调，保持播音员的专业克制。根据情境需要，你可以在任何你认为合适的地方添加任何非语言的自然声音——例如叹气声、笑声、咳嗽声、抽泣声等等。”,
”target_text”: ”根据国家表彰奖励工作决定，本次奖项设一等奖、二等奖、三等奖三个等级，获奖者将分别获颁荣誉奖章、证书及相应奖金，以表彰其在各自领域作出的突出贡献。”,
”expression_mode”: ”information_broadcast”,
”scene”: ”荣誉表彰公告”,
”subscene”: ”奖项等级与奖励内容说明”,
”situation”: ”角色在播报国家级奖项授予决定时需要说明奖项等级与奖励形式，并确保信息完整准确”,
”character_profile”: ”青年女性嗓音，音色明净纯正，口腔共鸣集中，吐字标准规范，整体气质平稳端方，具有新闻播音特有的专业与公信力”
}

C Gemini Case Study and Human Expressiveness Criterion

C.1 Complete Gemini output for one sample

The following nine outputs are copied verbatim from Gemini-3.6-Flash for test set `test_en_natural_language`, evaluated model `qwen3tts`.

Emotion Intensity

- **observed_behavior:** The speaker delivers the line with strong vocal projection, high vocal energy, and authoritative, urgent pitch contours. The delivery opens with an energetic, commanding call (”Everyone, please listen”) and maintains strong forward resonance, clear dynamic emphasis, and controlled emotional urgency throughout the passage.

- **supporting_evidence:** The vocal performance demonstrates clear cue convergence for an expansive, projected strategy: wide pitch inflections (particularly on ”listen” and ”shouts loudest”), firm articulatory weight on key content words (”not me”, ”sickest”, ”safe”, ”fair”), and high vocal energy with robust breath support suited for projecting across a room.

- **score_limiting_evidence:** Phonation stays clean, smooth, and well-controlled rather than pushing into extreme subglottic pressure, harsh glottal strain, creak, or dramatic pitch cracking. The performance maintains steady composure rather than escalating into maximum physiological tension or raw emotional breakdown.

- **rubric_application:** In the ‘role_expression‘ rubric, this performance displays distinct, robust emotional commitment and unambiguous dramatic stakes with prominent acoustic projection and weighted stress, fitting the High / Strong Emotional Intensity band (8–7).

- **score_rationale:** The vocal energy and authoritative emotional drive are consistently sustained across the full recording, making the emotional force strong and highly perceptible. Because it maintains polished vocal control without reaching maximal physical strain or extreme pitch peaks, a score of 7 accurately reflects its high, well-projected emotional intensity.

<−−EMOTION_INTENSITY_SCORE_START−−>
emotion_intensity_score: 7
<−−EMOTION_INTENSITY_SCORE_END−−>

Emotion Dynamic Naturalness

observed_behavior:
The speaker initiates the performance with a projected, attention-grabbing opening (”Everyone, please listen”), transitions smoothly into an authoritative and structured dynamic baseline (”Nurse Mbeki sets the order...”), and gradually modulates tone into an explanatory build before gently de-escalating to a reassuring, grounded conclusion (”Your turn will come; the list is fair”).

supporting_evidence:
Across the entire recording, shifts in pitch, intensity, tempo, and vocal effort move in tight acoustic coordination. Inter-phrase pauses function as realistic respiratory and cognitive resets rather than static silences. The dynamic arc moves naturally from firm crowd control to steady reassurance, with a convincing terminal decay in energy that reflects human physiological control.

score_limiting_evidence:
No material score-limiting evidence observed.

rubric_application:
The audio demonstrates flawless, organically integrated dynamic organization as described in the Band 9-10 rubric. Multi-parameter coordination across intensity, pitch trajectory, and phrase junctures ensures high fluid continuity without synthetic step-changes or unmotivated energetic spikes.

score_rationale:
The performance achieves a exceptionally natural emotional trajectory over time. The structural flow accurately reflects the temporal progression of a real speaker asserting order while reassuring a crowd, fully satisfying the highest criteria for emotional dynamic naturalness.

<−−EMOTION_DYNAMIC_NATURALNESS_SCORE_START−−>
emotion_dynamic_naturalness_score: 10
<−−EMOTION_DYNAMIC_NATURALNESS_SCORE_END−−>

Emotion Appropriateness

- **observed_behavior**: The speaker opens with a clear, projected call out to command attention across an open space, followed by an authoritative yet reassuring multi-phase trajectory. She establishes structural authority by deferring to the nurse with firm downward inflections on boundary lines (”not me”, ”not whoever shouts loudest”), transitions into a grounded explanatory tone, and concludes with a calming, definitive terminal fall.

- **supporting_evidence**:
- *Projection and Spatial Awareness*: The opening phrase (”Everyone, please listen”) is delivered with strong acoustic energy and elevated posture, effectively simulating projection across a noisy space like a gymnasium.
- *Dominance vs. Deference & Boundary Setting*: The stance shifts smoothly to project institutional order through crisp consonant articulation and stable, firm pitch contours during ”Nurse Mbeki sets the order, not me...”.
- *Multi-Phase Trajectory*: The delivery modulates smoothly from high-energy projection to reasoned explanation (”because that’s what keeps this whole shelter safe”) and settles into a steady, comforting downward cadence on ”Your turn will come; the list is fair.”

- **score_limiting_evidence**: No material score-limiting evidence observed.

- **rubric_application**: Under the *role_expression* rubric, this performance exemplifies top-tier contextual congruence (Score 10-9). The speaker fully integrates persona dynamics (bright, forward resonance), spatial requirements (gymnasium projection), and subtext (assertive rule enforcement paired with reassuring de-escalation).

- **score_rationale**: The recording flawlessly captures the character profile, scene constraints, and target text. The vocal stance dynamically adapts across functional phases, maintaining dramatic realism without slipping into generic anger or exaggerated theatricality.

<−−EMOTION_APPROPRIATENESS_SCORE_START−−>
emotion_appropriateness_score: 9
<−−EMOTION_APPROPRIATENESS_SCORE_END−−>

Prosody Richness

Observed behavior:
The recording exhibits continuous multi-dimensional prosodic modulation across pitch, dynamic vocal effort, speaking rate, and pause architecture. The speaker moves fluidly between different pitch registers, alters vocal weight across key focal points, varies local pacing from compressed delivery to deliberate lengthening, and uses clear structural pauses to delineate narrative units.

Supporting evidence:
- **Pitch Repertoire & Melodic Movement:** Wide pitch excursions and contour shifts occur across the entire utterance. High melodic peaks open the phrase (”Everyone”), followed by register resets on structural boundaries (”Nurse Mbeki”, ”not me”, ”shouts loudest”, ”sickest”).
- **Dynamic Effort Modulation:** Vocal intensity varies significantly, projecting strongly on focal target words (”order”, ”shouts loudest”, ”treated first”, ”whole shelter safe”) while dropping slightly on lower-intensity connective text.
- **Temporal Elasticity & Rate Regimes:** Local pacing alternates between rapid syllable bursts (”and not whoever shouts”) and expressive syllabic lengthening (”loudest”, ”safe”, ”fair”).
- **Phrasing & Pause Architecture:** The utterance is cleanly segmented by clear pauses and pitch resets after ”Everyone”, ”listen”, ”order”, ”me”, ”loudest”, and ”safe”, creating a diverse rhythmic cadence.

Score-limiting evidence:
No material score-limiting evidence observed.

Rubric application:
Under the ‘role_expression‘ rubric, this performance demonstrates expansive, continuous multi-parameter richness. Modulation across pitch contours, dynamic projection, temporal pacing, and phrase boundary architecture is sustained throughout the entire timeline, placing it firmly in the top tier (9-10).

Score rationale:
A score of 9 is fully justified as the performance exhibits continuous, multi-dimensional prosodic modulation throughout the entire utterance without collapsing into monotone delivery or static pacing.

<−−PROSODY_RICHNESS_SCORE_START−−>
prosody_richness_score: 9
<−−PROSODY_RICHNESS_SCORE_END−−>

Prosody Appropriateness

- **observed_behavior**: The speaker delivers the message with a firm, bright alto voice, sustained vocal projection, and a measured, authoritative cadence. Phrasing is broken by deliberate, meaningful micro-pauses that reinforce clarity and maintain control over the room.
- **supporting_evidence**: Contrastive and pragmatically critical words receive natural syllabic weight and pitch focus (e.g., ”Nurse Mbeki”, ”not *me*”, ”shouts *loudest*”, ”*sickest*”, ”*first*”, ”*fair*”). Downward terminal trajectories on ”not me” and ”the list is fair” effectively convey institutional resolve and finality, while maintaining enough warmth and steady rhythm to disarm and calm onlookers. Vocal effort accurately simulates public projection across a noisy shelter without collapsing into uncontrolled shouting.
- **score_limiting_evidence**: No material score-limiting evidence observed.
- **rubric_application**: The performance aligns with the Score 10-9 band (Exceptional Prosodic Appropriateness). Every suprasegmental feature—from pitch register baseline to temporal architecture and dynamic effort—converges cleanly to embody the character’s status and immediate scene goals.

- **score_rationale**: The execution fully satisfies all context requirements of the character profile and dramatic context with no audible prosodic flaws or unnatural inflections.

<−−PROSODY_APPROPRIATENESS_SCORE_START−−>
prosody_appropriateness_score: 10
<−−PROSODY_APPROPRIATENESS_SCORE_END−−>

Prosody Dynamic Naturalness

- **observed_behavior**: The speech performance presents a dynamic, loud, and projecting delivery appropriate for a public announcement. Phrase boundaries are distinctly marked by silences, but the transitions across phrase terminals, sustained vowels, and specific word endings exhibit noticeable synthetic rigidity, abrupt pitch quantization, and unnatural dynamic cut-offs.

- **supporting_evidence**: The macro-prosodic trajectory maintains a human-like directional energy contour across the utterance. The macro-phrasing reflects clear cognitive grouping (”Nurse Mbeki sets the order...”, ”The sickest are treated first...”), and phonatory onsets across major clause beginnings initiate with vocal fold adduction supported by high acoustic energy.

- **score_limiting_evidence**: Multiple phrase terminals and pitch transitions suffer from noticeable mechanical tendencies:
1. Phrase-final offsets on ”please listen” and ”this whole shelter safe” feature unnaturally clipped or abruptly gated dynamic tails without organic aerodynamic decay.
2. Pitch trajectories on sustained accented words such as ”shouts loudest” and ”treated first” exhibit discrete synthetic stepping/quantized pitch plateaus rather than fluid micro-prosodic inflections.
3. Pause intervals (e.g., between ”not me” and ”and not whoever...”) lack physiological respiratory integration, sounding uniformly placed and mechanically reset.

- **rubric_application**: According to the rubric for ‘role_expression‘, performances with noticeable mechanical stepping, rigid pacing, unnatural pitch plateaus, and premature clipped offsets fall within Band 6-5 (Moderate Dynamic Naturalness with Noticeable Mechanical Tendencies). The systemic presence of these artifacts across multiple phrase boundaries prevents placement in Band 7 or higher.

- **score_rationale**: A score of 5 is justified because while the overall broad dynamic trajectory carries clear expressive intent, recurring mechanical pitch stepping and gated phrase releases consistently disrupt the biomechanical continuity of the speech engine across the recording.

<−−PROSODY_DYNAMIC_NATURALNESS_SCORE_START−−>
prosody_dynamic_naturalness_score: 5
<−−PROSODY_DYNAMIC_NATURALNESS_SCORE_END−−>

Paralanguage Richness

Observed behavior:
The vocal performance is delivered with clear, steady articulation and projection, but features an extreme absence of audible non-lexical vocal events. Utterance onsets, inter-phrase syntactic transitions, and phrase terminations are cleanly gated without perceptible respiratory resets, friction-laden breath intakes, sighs, or vocal catches.

Supporting evidence:
Across the entire 13-second recording, major clausal and sentence boundaries—such as after ”please listen” (00:01), ”shouts loudest” (00:05), and ”shelter safe” (00:10)—consist entirely of silent pauses. There are no audible pre-speech inhalations framing the start of phrases, nor are there any audible exhalations or non-word vocalizations capping the ends of thoughts.

Score-limiting evidence:
The complete lack of audible respiratory resets, friction-laden gasps, vocalized catches, or terminal breath releases limits the temporal coverage of paralanguage to under 1% of the audio duration, preventing placement in any higher richness band.

Rubric application:
Under the *role_expression* rubric, this performance maps directly to the **2–1 (Trace to Zero Paralanguage)** anchor. Although the character profile suggests a high-exertion scene projecting across a crowded space, paralanguage richness evaluates solely the audible quantity and temporal coverage of non-lexical vocal events, which are virtually absent here.

Score rationale:
A score of 1 is assigned because non-lexical vocal events are completely inaudible across the recording, relying strictly on clean speech and silent pauses.

<−−PARALANGUAGE_RICHNESS_SCORE_START−−>
paralanguage_richness_score: 1
<−−PARALANGUAGE_RICHNESS_SCORE_END−−>

Paralanguage Diversity

- observed_behavior: The performance relies almost exclusively on clean lexical phonation separated by silent syntactic pauses. There is a complete absence of distinct non-lexical vocal events such as audible ingressive catch-breaths, egressive sighs, stance-marking scoffs, somatic gestures, or cognitive hesitations.
- supporting_evidence: Across key phrase transitions (e.g., at 00:01 after ”listen”, at 00:05 after ”loudest”, and at 00:09 after ”safe”), the boundaries contain silent syntactic pauses without any audible non-verbal vocalizations or respiratory events.
- score_limiting_evidence: The performance demonstrates no categorical breadth, completely foregoing audible non-lexical paralanguage features across all functional classes.
- rubric_application: In the role_expression mode, a performance that relies purely on unbroken lexical text and silent pauses falls squarely into the 1–2 score band (Minimal to absent category diversity).
- score_rationale: Because zero distinct non-lexical paralanguage categories are audibly realized throughout the recording, the category diversity reaches the absolute lower ceiling of the scale, justifying a score of 1.

<−−PARALANGUAGE_DIVERSITY_SCORE_START−−>
paralanguage_diversity_score: 1
<−−PARALANGUAGE_DIVERSITY_SCORE_END−−>

Paralanguage Appropriateness

Observed Behavior:
The speaker delivers the target text with disciplined respiratory control, utilizing clean, low-profile breath resets placed at key syntactic boundaries (e.g., prior to ”The sickest are treated first...”). The performance refrains from incorporating unmotivated sighs, theatrical gasps, or intrusive vocal fry.

Supporting Evidence:
Respiratory resets are executed quietly and naturally, supporting the vocal projection required for a crowd-calming directive without disrupting lexical intelligibility or speech flow. The clean pauses between rhetorical segments reinforce the character’s steady composure and institutional authority under high-stress conditions.

Score-Limiting Evidence:
No material score-limiting evidence observed.

Rubric Application:
Under the ‘role_expression‘ rubric, this delivery aligns with the Exemplary band (Score 9–10). The paralinguistic execution demonstrates strategic restraint and precise temporal placement. By avoiding melodramatic gasps or unmotivated non-lexical mannerisms, the vocalizations ground the character’s calm authority and dramatic credibility within the scene context.

Score Rationale:
A score of 10 is assigned because the non-lexical vocal events (and their controlled, solemn suppression) fit the character profile, dramatic stakes, and physical scenario seamlessly, demonstrating ideal paralinguistic parsimony and structural alignment.

<--PARALANGUAGE_APPROPRIATENESS_SCORE_START-->
paralanguage_appropriateness_score: 10
<--PARALANGUAGE_APPROPRIATENESS_SCORE_END-->

C.2 HUMAN EXPRESSIVENESS EVALUATION CRITERIA

Human evaluators compare two anonymized audios and choose the one that better realizes the intended expressive performance. The judgment considers the coordination of emotion, prosody, timing, pauses, vocal quality, and audible paralanguage across the complete utterance. A performance should be judged by perceptual clarity, contextual appropriateness, temporal coherence, and naturalness; variation or loudness alone is not sufficient evidence of better expressiveness.

The labels are 1 for Audio A preferred to Audio B, 2 for no clear difference, and 3 for Audio B preferred to Audio A. Evaluators listen to the complete pair before making a holistic judgment and base the decision on audible evidence rather than the written instruction alone.

Human Expressiveness Evaluation Criteria

Please treat the two audio clips as two renditions of the same character, the same scene, and the same target text. Judge them mainly by combining dimensions such as emotion, prosody, and paralanguage, focusing on comparing their overall differences in conveying expressive intent and in listening performance. During evaluation, first read the character description, scene information, and target text to clarify the speaker's identity, relationships, communicative purpose, emotional state, as well as the key points and changes in the text, and then make a judgment after listening to both audio clips in full.
Focus on the following aspects:
1. Whether the emotion and expressive intent match.
Judge whether the type, attitude, emotional coloring, and intensity of the emotion in the audio match the character, scene, communication purpose, and text content. Stronger emotion, greater volume, or more variation is not necessarily better; in scenes involving comfort, seriousness, fatigue, formality, or restraint, stable and restrained expression may be more appropriate than exaggerated performance.
2. Whether the emotion and expression develop naturally.
Observe whether the rises and falls, continuation, turns, and resolution of emotion throughout the entire passage are coherent. Note whether emotional changes have sufficient semantic and contextual basis, and whether local emphasis, emotional peaks, weakening, and endings connect naturally, avoiding sudden jumps, mechanical resets, unjustified emotional shifts, or disconnection between earlier and later expression.
3. Whether the prosody is rich and reasonable.
Comprehensively assess factors such as pitch, volume, vocal effort, speaking rate, rhythm, stress, syllable duration, phrasing, and pauses through listening. Good performance is not just the presence of variation, but also the use of appropriate values in appropriate places, forming clear layers according to text structure and expressive purpose. Prosodic variation should cover the entire passage, rather than appearing only once in a certain local part; continuously maintaining maximum volume, extreme pitch, or fast speaking rate does not equal rich prosody.
4. Whether pauses and paralanguage are meaningful and integrated into the expression.
Pay attention to audible non-lexical sounds such as breathing, sighing, light laughter, hesitation sounds, inhalation, humming, and throat clearing, and judge whether their type, quantity, variation, placement, and acoustic naturalness serve the current expression. Paralanguage is not better the more there is; restrained use or no use at all may also be a reasonable choice. Silent pauses, pitch changes, and persistent timbre characteristics cannot be directly treated as paralanguage events.
5. Whether the overall listening impression is unified, authentic, and organic.
Various acoustic cues should support one another; for example, emotion, pitch, volume, speaking rate, pauses, and paralanguage should form a consistent expressive direction. Note whether there are problems such as mechanicalness, stiffness, a spliced feeling, abrupt volume or pitch jumps, unnatural pauses, sudden cutoffs, and inconsistency between local performance and the overall context.

Table 17: Organization of the nine scoring dimensions.

Category	Sub-dimensions
Degree axis	Emotion Intensity; Prosody Richness; Paralanguage Richness; Paralanguage Diversity
Quality axis	Emotion Appropriateness; Emotion Dynamic Naturalness; Prosody Appropriateness; Prosody Dynamic Naturalness; Paralanguage Appropriateness

Human Expressiveness Evaluation Criteria (continued)

When comparing, the assessment should be based on the entire audio segment, while also paying attention to representative local positions. Issues that persist, have an obvious impact, or occur at key textual positions should be considered more important than minor issues that are occasional and do not affect overall comprehension. Do not judge an entire audio segment as better just because of one brief highlight, and do not reject the overall performance because of a single minor flaw.
Finally, make a pairwise choice:
- Audio1 win Audio2: audio1 has a relatively clear and more stable advantage in overall expression of intent, expression intensity, matching of emotion and scene, prosodic organization, dynamic naturalness, or use of paralanguage.
- Audio1 lose Audio2: audio2 has the more obvious overall advantage described above.
- Tie: the overall performance of the two audio clips is close, or their respective strengths and weaknesses offset each other, making it impossible to form a reliable overall preference.
The evaluation should be based on the evidence actually heard, and audio performance should not be inferred only from text, attribute descriptions, or model expectations; at the same time, try to distinguish pure audio quality, recording noise, pronunciation errors, or model identity preference from expressiveness itself.

D Additional sub-dimension-level Results

The main paper already includes the four English radar plots. The following figures in Figure 4 add four additional Chinese subsets.

E Additional Expression-mode-level Results

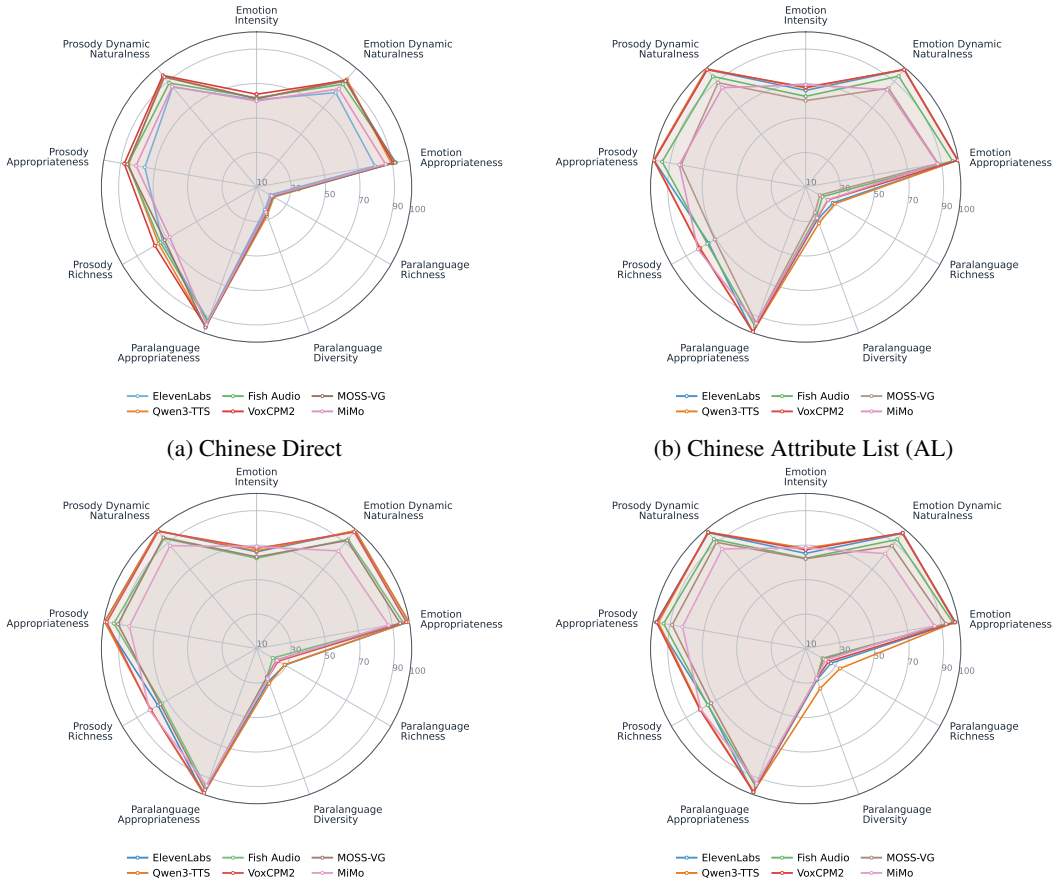
The main paper shows the results on all $8,000$ test samples in Figure 3. The following figures show the results of each subset in Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, Figure 11, and Figure 12, respectively.

F Evaluation Protocols and Prompts

We can divide all nine sub-dimensions into two categories (Table 17):

1. **Degree axis**: mainly reflects the degree of use of expressive resources, measuring indicators such as the quantity or intensity of expressive features, including Emotion Intensity, Prosody Richness, Paralanguage Richness, and Paralanguage Diversity.
2. **Quality axis**: dimensions related to expressive quality, including Emotion Appropriateness, Emotion Dynamic Naturalness, Prosody Appropriateness, Prosody Dynamic Naturalness, and Paralanguage Appropriateness.

Furthermore, in this section, we also show some expression-mode-specific scoring prompts used by the LALM evaluator. Due to limited space, we report some representative scoring prompts rather than presenting all 45 scoring prompts here. We have added all the detailed scoring prompts to the supplementary materials. In addition, we also provide the complete code and prompts for automated generation of expression-mode-specific scoring rubrics in the supplementary materials.



(a) Chinese Direct

(b) Chinese Attribute List (AL)

(c) Chinese Natural Language (NL)

(d) Chinese Scenario Performance Guidance (SPG)

Figure 4: Additional radar plots for the Chinese test subsets across the nine expressiveness dimensions. The English subsets are shown in Figure 2.

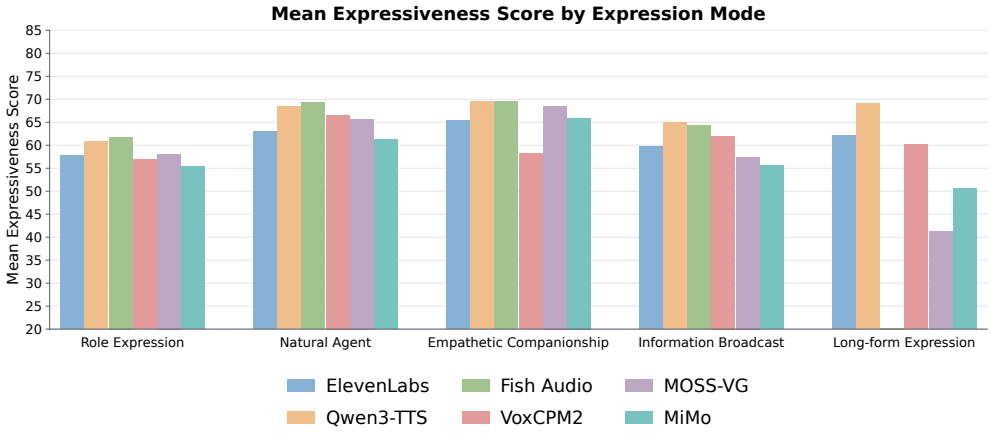


Figure 5: Mean expressiveness scores across five expression modes on the English Direct subset.

F.1 Role Expression

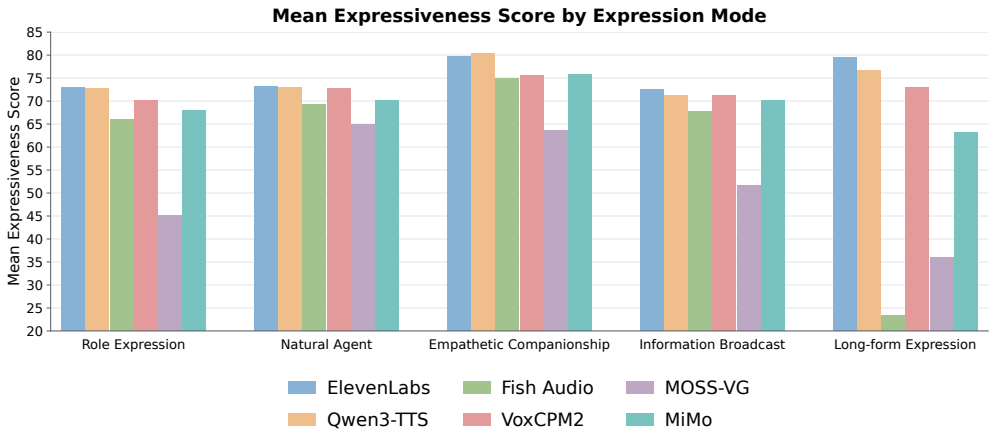


Figure 6: Mean expressiveness scores across five expression modes on the English Attribute List (AL) subset.

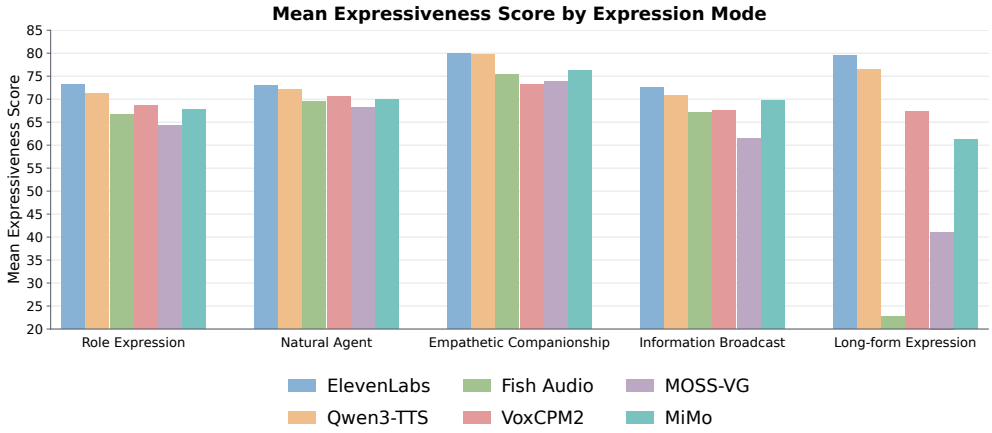


Figure 7: Mean expressiveness scores across five expression modes on the English Natural Language (NL) subset.

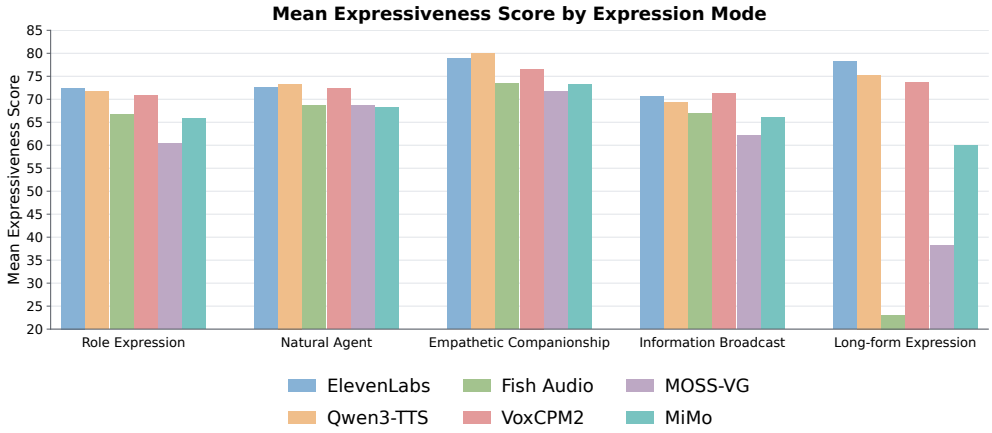


Figure 8: Mean expressiveness scores across five expression modes on the English Scenario Performance Guidance (SPG) subset.

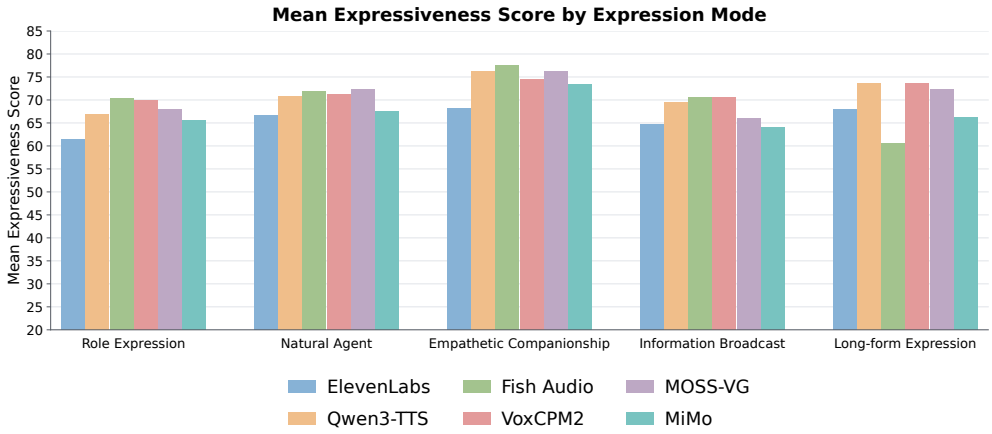


Figure 9: Mean expressiveness scores across five expression modes on the Chinese Direct subset.

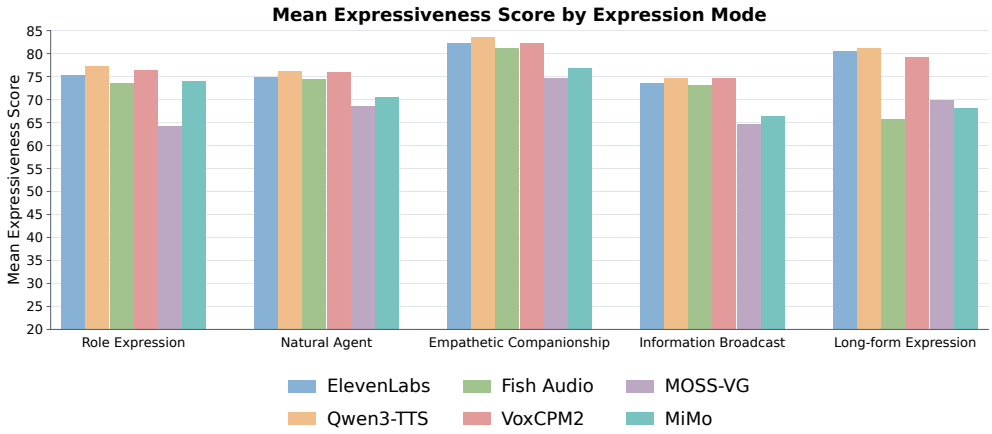


Figure 10: Mean expressiveness scores across five expression modes on the Chinese Attribute List (AL) subset.

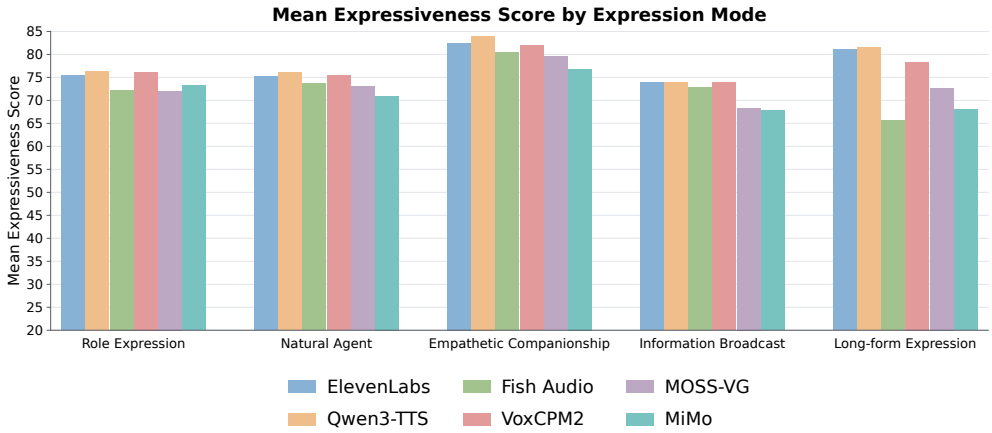


Figure 11: Mean expressiveness scores across five expression modes on the Chinese Natural Language (NL) subset.

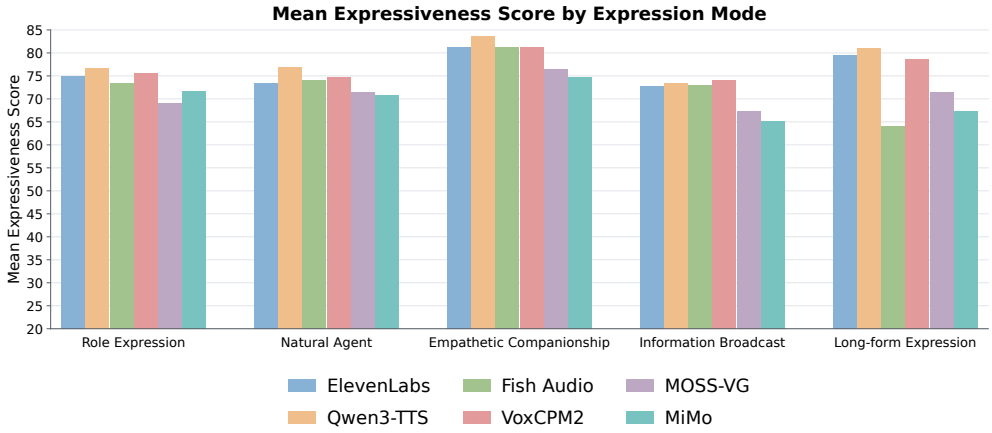


Figure 12: Mean expressiveness scores across five expression modes on the Chinese Scenario Performance Guidance (SPG) subset.

Scoring prompt by sub-dimension

Emotion Intensity

You are a professional evaluator of emotional intensity in speech.

You will receive one audio recording together with its intended character profile, scene, and target text. Evaluate only the target dimension defined below, using the expression-mode-specific rubric. Treat the reference context as the intended performance target, never as proof of what is audible.

Target expression mode:
role_expression (Role-based expressive performance)

Character profile:
{character_profile}

Scene:
{situation}

Target text:
{target_text}

Target dimension:
Emotional intensity (emotion_intensity)

Core definition:
The perceptual strength and salience of the emotion actually conveyed by the voice, as revealed by converging cues such as vocal energy, pitch, voice quality, articulation, tempo, pausing, breath, and tension.

Calibration rule:
Interpret high and low degree relative to the natural and plausible expressive range of the specified expression mode, not by one absolute acoustic scale shared across all modes. Meanwhile, a high degree does not mean that the behavior is contextually appropriate; appropriateness is evaluated separately.

Core boundary:
Assess strength rather than emotional appropriateness, temporal naturalness, or general acting quality. Loudness alone is not intensity, and restrained emotion may still be strongly perceptible through several coordinated cues.

Expression-mode-specific rubric:

Scoring prompt by sub-dimension (continued)

Expression-mode-specific understanding

In role-based expressive performance (role_expression), emotional intensity reflects the perceptual strength, salience, and physiological commitment of the emotion communicated through the voice. Unlike everyday neutral speech or factual reading, role-based speech encompasses a broad theatrical, narrative, and dramatic expressive range. This mode spans overt, explosive confrontational outbursts to deeply contained, high-stakes restraint, solemn proclamations, and vulnerable somatic entreaties.

Emotional intensity is calibrated relative to the natural and plausible expressive envelope of role performance. A high-intensity score in role_expression requires a substantial perceptual magnitude of emotional force within dramatic acting norms; this same acoustic magnitude in a casual conversational mode might register as anomalous or extreme.

Crucially, high emotional intensity in role performance manifests across two distinct vocal configurations:
1. **High-Acoustic-Power / Expansive Strategies:** Marked by elevated subglottic pressure, wide pitch excursions, pushed transglottal airflow, strained phonation, rapid clip, and resonant projection (e.g., active defiance, panicked cries, celebratory euphoria, or imperial decree).
2. **Low-Acoustic-Power / High-Tension Restrained Strategies:** Marked by dense medial compression, localized laryngeal constriction, steady vocal fry, unyielding pitch flatness, breath-damped voicing, prolonged tense pauses, or weighted consonantal bite (e.g., lethal suppressed rage, cold dread, suffocating grief, or desperate whispered entreaties).

Evaluators must recognize that raw decibel volume is not a proxy for intensity; tight, quiet physiological containment can project equal or greater emotional salience than an unconstrained shout.

Detailed evaluation criteria

Evaluators assess emotional intensity by observing the convergence and magnitude of acoustic, physiological, and articulatory cues across the utterance:

1. **Pre-phonatory and Respiratory Cues:**
- *High Intensity:* Audible physiological preparatory markers establishing immediate emotional stakes, such as sharp inhalatory gasps, trembling vocalized breaths, dry nasal scoffs, deep sighed exhalations, or low chuckled onsets.
- *Low Intensity:* Silent, passive, or purely functional vegetative breathing with no affective coloring.

2. **Phonation, Voice Quality, and Glottal Tension:**
- *High Intensity:* Strong vocal tract activation—manifesting as acoustic strain, hyperfunctional medial compression, creak/fry under tension, marked breathy exhaustion, or resonant acoustic projection.
- *Low Intensity:* Modal, neutral phonation with minimal vocal fold tension, lax glottal engagement, and flat vocal resonance.

3. **Pitch Range, Contours, and Dynamic Modulation:**
- *High Intensity:* Dynamic escalation toward climactic peaks, expansive pitch ceilings, abrupt register breaks, sweeping melodramatic cadences, or conversely, an unwavering, rigidly sustained low-register monotone conveying acute suppressed menace.
- *Low Intensity:* Narrow, typical speaking fundamental frequency range without significant pitch inflection or intentional dramatic flattening.

4. **Temporal Dynamics, Articulatory Weighting, and Pausing:**
- *High Intensity:* Deliberate consonantal bite, extended vowel duration under stress, marked rhythmic deceleration or agitated acceleration, and charged, breath-held inter-phrase pauses.
- *Low Intensity:* Standard, unaccented conversational pacing, uniform syllable durations, and casual, unweighted articulation.

5. **Temporal Distribution and Cue Convergence:**
- Evaluators must weigh the cumulative presence and coordinated alignment of multiple cues across the utterance. Sustained macro-level intensity or well-developed dynamic escalations to powerful focal peaks score higher than isolated, unintegrated micro-cues.

Scoring prompt by sub-dimension (continued)

Evaluation procedure

1. **Acoustic and Physiological Cue Inventory:**
Listen to the performance purely for vocal and respiratory behavior. Identify specific markers: respiratory onsets, phonatory strain, breath damping, pitch excursions, pause tension, and articulatory weighting.
2. **Expressive Strategy Identification:**
Determine whether the performance utilizes an expansive (overt acoustic power) or a contained/restrained (low-volume, high-tension) strategy. Evaluate cue strength according to that strategy's specific physiological markers.
3. **Synthesis of Cue Convergence and Dynamic Arc:**
Assess how consistently and intensely these cues coordinate. Does the emotional force peak at key dramatic turns? Is high tension maintained across phrase boundaries?
4. **Boundary Decoupling Check:**
Consciously separate the perceived *strength* of the emotion from:
- *Appropriateness:* Do not downgrade a performance for being overacted or out-of-context, nor upgrade a weak delivery because the script called for subtle acting.
- *Acting Quality/Aesthetic Conviction:* Evaluate raw emotional strength regardless of whether the dramatic interpretation feels convincing or artistically refined.
- *Loudness Bias:* Verify that quiet, hushed, or suppressed deliveries with heavy vocal tension are not underscored due to low decibels.
5. **Scale Assignment:**
Assign a score within the 1-10 bands based on where the perceptual magnitude falls along the role_expression mode's plausible expressive continuum.

Boundaries and common rating errors

- **Conflation with Acoustic Loudness (Volume Bias):** Loudness is an acoustic amplitude measure, whereas intensity is an integrated perceptual assessment. A relaxed yell has lower emotional intensity than an icy, throat-constricted threat delivered in a low register.
- **Conflation with Emotional Appropriateness / Contextual Fit:** Rating how "right" the emotion is for the narrative is strictly prohibited. An exaggerated, melodramatic scream in a quiet library scene must receive a high intensity score even if it is contextually absurd.
- **Conflation with Acting Quality / Believability:** A performance can feel artificial, forced, or melodramatic while still exhibiting high emotional intensity. Conversely, an expertly subtle, believable performance may have deliberately low emotional intensity.
- **Overlooking Restrained or Vulnerable Intensity:** Failing to award high scores to quiet, breath-damped pleas, cold suppressed fury, or grief-stricken vocal fry due to the absence of shouting.
- **Single-Cue Fixation:** Basing a high or low score entirely on an isolated gasp, brief pitch crack, or single loud word without evaluating the broader acoustic contour and converging cue matrix.

Expression-mode-calibrated score anchors

- **10-9: Maximal / Extreme Emotional Intensity**
- The voice projects extraordinary, overwhelming emotional force, fully utilizing the outer boundaries of role-based expressive potential.
- Characterized by profound cue convergence: explosive subglottic pressure with extreme pitch excursions, severe phonatory strain, or deeply physicalized, breath-held containment with intense laryngeal constriction, shuddering onsets, and hyper-articulated consonantal releases.
- The emotional charge saturates the entire delivery or drives an unmistakable, powerful climactic surge.

- **8-7: High / Strong Emotional Intensity**
- The performance exhibits distinct, robust emotional commitment and high perceptual salience, clearly above standard conversational or narrative levels.
- Manifests through evident cue coordination: noticeable pitch widening, prominent vocal fry/creak under tension, pushed or breath-damped airflow, weighted articulatory stress, or audibly charged pre-phonatory breathing.
- The delivery conveys unambiguous dramatic stakes and emotional drive across major phrases.

Scoring prompt by sub-dimension (continued)

- **6-5: Moderate / Intermediate Emotional Intensity**
- Demonstrates clear, audible emotional engagement within a conventional, balanced dramatic range.
- Cues are noticeable but moderate: standard expressive pitch inflections, mild vocal tract tension, occasional breath support variations, and measured tempo shifts without extreme physiological strain or deep suppressed containment.
- The vocal delivery is clearly expressive rather than flat, but lacks high acoustic power or intense physiological pressure.

- **4-3: Low / Subdued Emotional Intensity**
- The voice exhibits only mild or faint emotional coloring, remaining close to a detached or casual baseline.
- Acoustic cues are sparse or weak: minimal pitch variation, lax phonation, conventional pacing, unweighted consonants, and an absence of dramatic breath markers or laryngeal tension.
- The character's emotional state registers as placid, minimally invested, or emotionally distant.

- **2-1: Minimal / Negligible Emotional Intensity**
- Complete or near-complete absence of perceived emotional force; the delivery is entirely flat, inert, or mechanical.
- No audible physiological commitment: uniform pitch, relaxed vocal folds, unmodulated tempo, and silent, purely functional respiration.
- Represents the extreme floor of emotional expression within role performance, conveying zero dramatic urgency or affective weight.

Notes:
1. Listen to the entire recording before deciding on a score.
2. Identify behavior that is actually audible and localize important evidence where possible.
3. Use the character profile, scene, and target text only to interpret the intended performance and any context-dependent criteria.
4. Apply the expression-mode-specific rubric and calibration consistently while respecting the dimension boundary.
5. Base the final score on the whole utterance and weigh isolated moments in proportion to their salience and coverage.
6. Actively check for score-limiting audible evidence instead of defaulting to a top-band score because the broad intent appears plausible or no gross mismatch is obvious. Do not invent limitations that are not audible.

For this degree axis, score_limiting_evidence means audible evidence that limits placement at a higher degree on the scale. It is not necessarily a quality defect.

Before the score block, provide a concise English report containing:
- observed_behavior: the relevant behaviors heard for this dimension
- supporting_evidence: concrete acoustic and temporal evidence supporting the assigned level or quality
- score_limiting_evidence: concrete audible evidence limiting a higher score; write "No material score-limiting evidence observed" only when that is genuinely supported by the full recording
- rubric_application: how the evidence maps to the expression-mode-specific criteria
- score_rationale: why the selected score band and integer are justified

At the very end, output exactly this block with one integer from 1 to 10 and no extra text inside it:
<--EMOTION_INTENSITY_SCORE_START-->
emotion_intensity_score: [1-10]
<--EMOTION_INTENSITY_SCORE_END-->

Paralanguage Richness

You are a professional evaluator of paralanguage richness in speech.

You will receive one audio recording together with its intended character profile, scene, and target text. Evaluate only the target dimension defined below, using the expression-mode-specific rubric. Treat the reference context as the intended performance target, never as proof of what is audible.

Scoring prompt by sub-dimension (continued)

Target expression mode:
role_expression (Role-based expressive performance)

Character profile:
{character_profile}

Scene:
{situation}

Target text:
{target_text}

Target dimension:
Paralanguage richness (paralanguage_richness)

Core definition:
The density, salience, temporal coverage, and expressive contribution of clearly audible non-lexical vocal events, such as breaths, sighs, laughs, sobs, gasps, hesitation sounds, or other non-word vocalizations.

Calibration rule:
Interpret high and low degree relative to the natural and plausible expressive range of the specified expression mode, not by one absolute acoustic scale shared across all modes. Meanwhile, a high degree does not mean that the behavior is contextually appropriate; appropriateness is evaluated separately.

Core boundary:
Assess the amount and coverage of audible non-lexical events rather than category variety or contextual appropriateness. Ordinary pitch movement, lexical stress, tempo, and silent pauses are not paralanguage events.

Expression-mode-specific rubric:
Expression-mode-specific understanding

In role-based expressive performance (role_expression), paralanguage richness reflects the degree to which a performer populates the vocal track with clearly audible, non-lexical physiological and vocal events—such as ingressive breath intakes, friction-laden gasps, sighs, groans, chuckles, scoffs, sobbing catches, and exhalatory releases. Within this expressive mode, characters span a wide spectrum from extreme physiological exertion and emotional volatility to highly controlled, ritualized, or stoic delivery.

Paralanguage in role_expression characteristically manifests through:
- **Pre-speech preparatory onsets**: Audible breath intakes, sharp gasps, or voiced sighed releases occurring prior to initial phonation that establish physical readiness, emotional tension, or conversational stance.
- **Inter-phrase syntactic segmentation**: Audible inspiratory breath resets populating clausal and grammatical junctures, maintaining continuity and narrative pacing.
- **Terminal and transitional releases**: Audible exhalations, chest-resonated sighed offsets, and breathy dissipations capping thoughts or marking rhetorical transitions.
- **Clausal pivots and cognitive hesitations**: Sudden inspiratory catches, glottalized hesitations, or friction surges directly preceding dramatic turns, accusations, or real-time calculations.
- **Context-conditioned variants**: High-exertion or distress environments manifest near-continuous labored breathing; celebratory or informal scenes incorporate voiced laugh bursts and chuckles; dismissive interactions feature nasalized scoffs and dry snorts.
- **Restrained strategies**: Formal, authoritative, or clinically detached roles utilize low-amplitude micro-inhalations or cleanly gated silences, where paralanguage coverage is deliberately minimal.

Because role_expression permits extreme dramatic ranges (from total respiratory suppression to heavy physiological saturation), an evaluation of richness must be calibrated to the mode's plausible expressive envelope. A score of 8 in role_expression represents a density and acoustic prominence appropriate to highly animated or physically demanding role acting, which would constitute an extreme outlier if observed in formal read-aloud speech.

Scoring prompt by sub-dimension (continued)

Detailed evaluation criteria

Evaluators must assess the cumulative density, acoustic salience, temporal coverage, and structural placement of non-lexical vocal events across the performance:

1. **Event Density (Frequency of Occurrence)**:
- Evaluates how often non-lexical events occur per unit time or syntactic unit.
- High density features events at nearly every grammatical boundary, pre-speech onset, and utterance tail. Low density features isolated or entirely absent events.

2. **Temporal Coverage (Duration Occupied)**:
- Evaluates the proportion of total performance time occupied by non-lexical events.
- High coverage includes sustained laugh tails, extended sighed exhalations, and labored respiratory resets (e.g., 15% to 25% or more of total audio duration). Low coverage is confined to brief micro-events or zero coverage (¡1%).

3. **Acoustic Salience and Prominence**:
- Evaluates the energy, audibility, and acoustic definition of the events relative to lexical speech and background noise.
- Highly salient events exhibit clear frictional turbulence, distinct voicing, sharp glottal catches, or prominent resonance. Low salience events are quiet, soft, low-friction micro-inhalations hovering near the ambient floor.

4. **Structural Integration and Distribution**:
- Evaluates how non-lexical events are distributed across the utterance structure (pre-speech framing, mid-phrase clausal pivots, inter-phrase resets, and terminal decay).
- High richness displays multi-locus distribution across onsets, pauses, and offsets; low richness is strictly localized or absent.

Weighing Converging Cues**: Evaluators must synthesize density and temporal coverage as primary drivers, with acoustic salience as a scaling factor. A high frequency of barely audible micro-breaths receives a moderate score, whereas frequent, sustained, and highly audible respiratory or voiced non-lexical events achieve the highest score tiers.

Evaluation procedure

1. **Acoustic Mapping and Identification**: Listen to the complete recording, identifying and cataloging all audible non-lexical vocal events (e.g., pre-speech breaths, clausal inhalations, sighs, scoffs, laughs, gasps, throat-clearing, voiced groans).
2. **Temporal Coverage and Salience Estimation**: Estimate the cumulative time occupied by these non-lexical events relative to total speech duration and assess their acoustic prominence (prominent, moderate, or subtle/whispered).
3. **Distributional Analysis**: Map whether non-lexical events occur across multiple structural positions (framing utterance onsets/offsets, populating phrase boundaries, interrupting clausal pivots) or are restricted to rare, isolated points.
4. **Mode-Calibrated Degree Assignment**: Match the observed density, duration, and salience against the role_expression calibrated scale (ranging from complete gated silence to saturated respiratory coverage).
5. **Context and Quality Decoupling**: Ensure the score reflects solely the physical quantity, salience, and temporal coverage of non-lexical events. Do not penalize or reward the score based on whether the non-lexical events feel emotionally appropriate, dramatic, or natural for the character.

Boundaries and common rating errors

- **Appropriateness vs. Richness Confusion**: Do not lower the score because frequent sobbing, gasping, or laughing feels overacted, exaggerated, or ill-suited to the text; similarly, do not award a high score to a stoic, silent performance merely because suppression fits the character. Score purely the observable presence and coverage of paralanguage.

Scoring prompt by sub-dimension (continued)

- **Conflation with Lexical Prosody**: Pitch excursions, intonational contouring, lexical stress, volume dynamics, and vowel lengthening on actual words are prosodic features, not paralanguage events. Only non-lexical vocalizations count.
- **Conflation with Silent Pauses**: Unvoiced, clean gaps in the acoustic stream are silent pauses, not paralanguage. A dramatic two-second silent pause contributes zero paralanguage coverage unless accompanied by audible breathing, friction, or vocal catches.
- **Conflation with Phonologized Fillers**: Formal lexical hesitations (e.g., standard lexicalized words) must be distinguished from pure non-lexical vocalizations (e.g., open-throat grunts, unphonologized hesitation catches, sighed glottal stops).
- **Single-Cue Bias**: Do not assign a top-tier score based on a single isolated multi-second laugh or one loud pre-speech gasp if the remainder of the performance contains zero paralanguage. Evaluate cumulative temporal coverage and density across the full performance.
- **Recording Artifact Misattribution**: Ensure microphone pops, room rumble, or technical noise gates are not mistaken for organic non-lexical vocal events.

Expression-mode-calibrated score anchors

- **10-9 (Saturated / Pervasive Paralanguage)**: Non-lexical vocal events saturate the performance, occupying substantial temporal coverage (typically 15% to 25%+ of total audio duration). Audible respiratory onsets precede speech, prominent friction-laden or voiced breath resets populate nearly every clausal boundary, and terminal sighed releases, gasps, or laughter tails frame phrase endings with high acoustic prominence.
- **8-7 (High Richness / Frequent Prominent Events)**: Frequent, clearly audible non-lexical vocalizations across most major syntactic boundaries, onsets, or climactic pivots (roughly 8% to 15% temporal coverage). Breaths, sighs, scoffs, or chuckles are distinctly audible, well-defined above the ambient floor, and consistently contribute to the vocal texture throughout the delivery.
- **6-5 (Moderate Richness / Regular Inter-Phrase Pacing)**: Non-lexical events occur regularly at major grammatical pauses and select onsets (roughly 4% to 8% temporal coverage). Inhalations and exhalations are clearly audible upon listening but exhibit moderate duration and salience, serving primarily as steady inter-phrase segmentation without heavy friction, dramatic gasps, or extended voiced tails.
- **4-3 (Sparse / Subdued Paralanguage)**: Non-lexical events are infrequent, highly localized, or acoustically subdued (roughly 1% to 3% temporal coverage). Manifests as brief, low-amplitude micro-inhalations near the noise floor or a single isolated sigh/gasp across an extended performance, leaving the vast majority of phrasal transitions quiet.
- **2-1 (Trace to Zero Paralanguage)**: Paralanguage events are virtually absent or completely inaudible (less than 1% coverage to absolute zero). Utterance onsets, clausal boundaries, and offsets are cleanly gated or accompanied only by silent pauses, with no detectable audible breaths, gasps, sighs, laughs, or non-lexical vocalizations.

Notes:
1. Listen to the entire recording before deciding on a score.
2. Identify behavior that is actually audible and localize important evidence where possible.
3. Use the character profile, scene, and target text only to interpret the intended performance and any context-dependent criteria.
4. Apply the expression-mode-specific rubric and calibration consistently while respecting the dimension boundary.
5. Base the final score on the whole utterance and weigh isolated moments in proportion to their salience and coverage.
6. Actively check for score-limiting audible evidence instead of defaulting to a top-band score because the broad intent appears plausible or no gross mismatch is obvious. Do not invent limitations that are not audible.

For this degree axis, score_limiting_evidence means audible evidence that limits placement at a higher degree on the scale. It is not necessarily a quality defect.

Before the score block, provide a concise English report containing:
- observed_behavior: the relevant behaviors heard for this dimension
- supporting_evidence: concrete acoustic and temporal evidence supporting the assigned level or quality
- score_limiting_evidence: concrete audible evidence limiting a higher score; write "No material score-limiting evidence observed" only when that is genuinely supported by the full recording

Scoring prompt by sub-dimension (continued)

- rubric_application: how the evidence maps to the expression-mode-specific criteria
- score_rationale: why the selected score band and integer are justified

At the very end, output exactly this block with one integer from 1 to 10 and no extra text inside it:
<−−PARALANGUAGE_RICHNESS_SCORE_START−−>
paralanguage_richness_score: [1-10]
<−−PARALANGUAGE_RICHNESS_SCORE_END−−>

Prosody Appropriateness

You are a professional evaluator of prosodic appropriateness in speech.

You will receive one audio recording together with its intended character profile, scene, and target text. Evaluate only the target dimension defined below, using the expression-mode-specific rubric. Treat the reference context as the intended performance target, never as proof of what is audible.

Target expression mode:
role_expression (Role-based expressive performance)

Character profile:
{character_profile}

Scene:
{situation}

Target text:
{target_text}

Target dimension:
Prosodic appropriateness (prosody_appropriateness)

Core definition:
How well both moment-to-moment values and trajectories of pitch register and range, perceived vocal effort and loudness, speaking rate, rhythm, phrasing, and pauses fit the character, scene, target text, local emphasis, and communicative intent.

Calibration rule:
Evaluate this quality dimension according to its own definition. Do not replace it with the amount of expression present.

Core boundary:
Assess the contextual fit of both levels and changes rather than sheer amount or transition smoothness. Interpret pitch relative to the speaker and loudness as perceived vocal effort or projection, not absolute hertz, decibels, or recording gain.

Expression-mode-specific rubric:
Expression-mode-specific understanding

Prosodic appropriateness within role-based expressive performance measures how accurately and organically the acoustic properties of speech—pitch register and range, perceived vocal effort, speaking rate, rhythm, phrasing, and pause structures—align with the speaker's portrayed persona, interpersonal status, immediate environment, dramatic objectives, and underlying textual semantics.

In dramatic and narrative performance, prosody is not a static decorative layer; it is the primary vehicle for signaling character dynamics, tactical shifts, and emotional subtext. High prosodic appropriateness is demonstrated when prosodic choices reflect both moment-to-moment textual nuances and broader scene-level trajectories:

- **Recurrent Prosodic Realizations:**

Scoring prompt by sub-dimension (continued)

- *Structural and Evidentiary Punctuation:* Deceleration of articulation rate, elongated syllabic duration, and distinct micro-pauses or caesuras when presenting key evidence, listing critical conditions, or isolating pivotal arguments, giving weight to information without relying solely on increased volume.
- *Authoritative and Resolute Cadences:* Grounded terminal downward pitch trajectories, heavy syllabic weight, and lower pitch register baselines to communicate finality, legal or military authority, firm refusals, or institutional resolve.
- *Duty and Persona Pivots:* Fluid modulation across interpersonal warmth and professional duty, transitioning from resonant, flexible contours to flattened, crisp, and metered delivery when institutional boundaries, audits, or hard realities are asserted.
- *High-Stakes Escalation:* Syllabic compression, elevated pitch register, and heightened vocal projection deployed during acute confrontation, panic, or urgent self-defense, resolving into grounded baselines once control is regained.

- **Context-Conditioned Alternatives:**
- *Authority vs. Subordination:* Senior authority may manifest as unhurried, measured tempos, resonant low-to-mid registers, and deliberative silences, or alternatively as sharp, elevated public decrees. Subordinate or deferent stances leverage higher pitch onsets, breath-infused phonation, rapid explanatory bursts, and soft trailing falls.
- *Tactful De-escalation vs. Confrontational Dominance:* De-escalation utilizes warm chest resonance, gentle trailing falls, and reduced vocal effort to disarm; confrontational dominance uses staccato rhythms, punchy syllable onsets, abrupt pitch peaks, and hard projection to enforce pressure.
- *Solemnity vs. Camaraderie:* Solemn, formal, or high-risk settings favor hushed projection, dry articulation, restricted pitch range, and strictly metered pacing; informal or communal environments thrive on buoyant pitch excursions, audible breath laughter, and dynamic rhythmic pulsing.

- **Restrained and Sparse Strategies:**
- Appropriateness frequently requires acoustic minimalism rather than dynamic acrobatics. In intimate, nocturnal, or life-threatening crisis contexts, suppressed vocal effort, low-register anchoring, and breath-dominated phonation convey urgency or physical exhaustion far more credibly than overt shouting. Similarly, analytical detachment or cold defiance is authentically realized through regularized pacing, uniform pitch contours, and unyielding emotional flatness.

Detailed evaluation criteria

Evaluators must assess prosodic appropriateness through the convergence of multiple acoustic cues across three core dimensions:

1. **Pitch Register, Range, and Contour Trajectories:**
- *Baseline Appropriateness:* Does the speaker's baseline pitch register fit the character's physical state, age, social standing, and emotional context relative to the speaker's natural vocal anatomy?
- *Contour Modulation:* Do pitch inflections support the communicative intent (e.g., terminal falls for definitive commands, rising or suspended inflections for genuine inquiry or uncertainty, abrupt peaks for sudden shock or outrage)?
- *Dynamic Register Shifts:* Are transitions between vocal registers motivated by dramatic shifts (e.g., switching from personal empathy to cold institutional firmness) rather than arbitrary vocal instability?

2. **Perceived Vocal Effort, Projection, and Dynamic Weight:**
- *Effort Alignment:* Is vocal projection calibrated to the acoustic space and interpersonal distance (e.g., hushed intimacy in close quarters versus projected resonance in command or public address)?
- *Stress and Focus:* Is perceived loudness and syllabic weight placed on pragmatically operative words, contrastive elements, and key narrative pivots, rather than landing arbitrarily on unstressed functional words?
- *Restraint vs. Escalation:* Does the speaker modulate vocal energy to reflect internal tension, suppression, or explosive release in accordance with the scene's stakes?

3. **Temporal Architecture: Rate, Rhythm, Phrasing, and Pauses:**
- *Rate Dynamics:* Does articulation speed accelerate or decelerate logically with emotional arousal, cognitive processing, deliberate rhetoric, or physical fatigue?
- *Rhythmic Structuring:* Does the rhythm reflect the character's mental state—ranging from smooth, regularized meter in analytical control to broken, staccato bursts during panic or confrontation?

Scoring prompt by sub-dimension (continued)

- *Phrasing and Syntactic Cohesion:* Do breath groups and phrase boundaries reinforce semantic clarity, dramatic suspense, or tactical maneuvering?
- *Pause Intentionality:* Are pauses, micro-caesuras, and silences timed to represent real-time thought, emotional weight, non-verbal calculation, or conversational turn-taking, rather than appearing as accidental hesitancy or mechanical pacing?

4. **Integration, Temporal Coverage, and Weighting:**
- *Global Trajectory vs. Local Nuance:* Evaluators must weigh sustained scene-level consistency alongside moment-to-moment responsiveness. A performance that maintains overall tone but fails on critical semantic pivots is flawed; conversely, localized brilliant moments cannot compensate for fundamentally misjudged baseline prosody.
- *Convergence of Cues:* High ratings require that pitch, energy, and timing converge to support a single communicative objective. Incongruent combinations (e.g., shouting with an unhurried, relaxed cadence during an urgent crisis) indicate poor prosodic control.

Evaluation procedure

1. **Context and Intent Formulation:**
- Analyze the target text, dramatic scene, character persona, power dynamic, and immediate physical or interpersonal environment prior to evaluating acoustic features.
- Establish the baseline dramatic requirements: What is the character's objective, emotional state, status relative to the listener, and environmental constraint?

2. **First Acoustic Pass (Global Architecture & Baseline Calibration):**
- Listen to the full performance to establish the speaker's baseline pitch register, perceived vocal effort, and default tempo relative to their vocal apparatus.
- Assess whether the overall prosodic profile matches the macroscopic requirements of the persona and scene (e.g., authoritative command, subdued confession, heated argument).

3. **Second Acoustic Pass (Local Modulation & Convergence Analysis):**
- Track local prosodic movements against semantic markers:
- Check evidentiary and structural punctuation (e.g., syllabic elongation, rate deceleration on key points).
- Check terminal cadences (e.g., authoritative downward glides versus unnatural upward drift).
- Check register shifts and vocal effort changes at emotional or tactical pivot points.
- Evaluate pause placement and duration for cognitive and dramatic legitimacy.

4. **Third Acoustic Pass (Discrepancy and Artifact Verification):**
- Identify any moments of prosodic mismatch: unmotivated pitch spikes, misplaced lexical emphasis, arbitrary tempo fluctuations, rhythmically unnatural pauses, or incongruent vocal effort.
- Verify whether restrained or flat delivery reflects deliberate, appropriate dramatic choice (e.g., detached calculation, suppressed grief) or generic, disengaged reading.

5. **Synthesis and Score Assignment:**
- Synthesize the balance between macro-level appropriateness and micro-level precision.
- Match the observed evidence against the calibrated score bands to assign the final rating.

Boundaries and common rating errors

- **Prosodic Appropriateness vs. Prosodic Variety / Amount of Expression:**
- *Error:* Rewarding exaggerated pitch acrobatics, wide dynamic swings, or theatrical over-inflection in scenes that call for subdued, monotone, or restrained delivery.
- *Boundary Rule:* Appropriate prosody can be entirely flat, quiet, or rhythmically uniform if the role demands cold detachment, exhaustion, or stealth. Evaluate whether the prosody fits the context, not how much prosodic movement is displayed.

- **Prosodic Appropriateness vs. Acoustic Recording Quality / Decibel Gain:**
- *Error:* Conflating high recording gain or absolute microphone volume with authoritative vocal projection, or penalizing low-gain recordings as lacking vocal energy.

Scoring prompt by sub-dimension (continued)

- *Boundary Rule:* Evaluate *perceived vocal effort* and physiological projection relative to the character's context, independent of electronic amplification or mastering levels.

- **Prosodic Appropriateness vs. Fluency / Naturalness:**
- *Error:* Penalizing deliberate dysfluencies, hesitation pauses, broken rhythms, or abrupt pitch breaks that are contextually necessary for portraying panic, trauma, cognitive strain, or physical injury.
- *Boundary Rule:* Fluency is smooth execution; appropriateness is contextual alignment. In high-stakes role performance, broken or jagged prosody is often the most appropriate realization.

- **Prosodic Appropriateness vs. Pronunciation and Articulation Quality:**
- *Error:* Conflating crisp consonant articulation or phonemic precision with prosodic fit.
- *Boundary Rule:* Articulation concerns phonetic clarity; prosody concerns suprasegmental contours of pitch, time, and effort across phrases and utterances.

- **Stereotype and Archetype Conflation:**
- *Error:* Assuming that authority must always be deep and slow, or that excitement must always be high-pitched and fast.
- *Boundary Rule:* Accept diverse valid prosodic realizations that credibly fulfill the communicative intent within the specific context.

Expression-mode-calibrated score anchors

Score 10-9: Exceptional Prosodic Appropriateness
- **Pitch & Register:** Pitch register, range, and intonation trajectories align impeccably with the character persona, interpersonal power balance, and narrative subtext. Contour changes precisely match semantic intent, deploying downward terminal cadences for resolve and calculated register shifts during tactical pivots.
- **Effort & Dynamics:** Perceived vocal effort and dynamic projection perfectly mirror the acoustic environment and dramatic stakes, effortlessly shifting between subtle, suppressed tension and fully projected authority without artificiality.
- **Timing & Phrasing:** Pacing, rhythmic structure, and pause durations demonstrate complete mastery. Pauses function convincingly as cognitive processing or dramatic leverage; evidentiary points and structural pivots are effortlessly highlighted through tempo modulation and syllabic weighting.
- **Holistic Integration:** Every suprasegmental cue converges seamlessly; restrained or explosive deliveries are entirely motivated by context.

Score 8-7: Good Prosodic Appropriateness
- **Pitch & Register:** Pitch register and contours generally fit the character, scene, and communicative intent well. Key emotional and tactical shifts are supported by appropriate pitch adjustments, with only minor, isolated instances of neutral or slightly predictable intonation.
- **Effort & Dynamics:** Vocal effort and loudness are well-calibrated to the setting and interpersonal dynamic. Stress is consistently placed on operative words and dramatic pivots, with rare moments of slight over- or under-projection.
- **Timing & Phrasing:** Speaking rate and phrasing support semantic clarity and dramatic momentum. Pauses are generally purposeful and natural, and tempo shifts appropriately distinguish narrative or evidentiary emphasis with only minor lapses in timing.
- **Holistic Integration:** Prosodic features consistently reinforce the scene's emotional logic, with few minor inconsistencies that do not disrupt overall credibility.

Score 6-5: Moderate / Adequate Prosodic Appropriateness
- **Pitch & Register:** Pitch register is acceptable for the broad archetype but lacks nuance. Intonation contours convey basic communicative intent (e.g., questions versus statements) but may sound generic, slightly theatrical, or occasionally unaligned with subtle subtext.
- **Effort & Dynamics:** Perceived vocal effort is functional but occasionally misjudged—such as maintaining an identical projection level across both intimate subtext and heightened conflict, or placing sporadic emphasis on non-critical words.
- **Timing & Phrasing:** Articulation rate is largely steady but shows limited contextual flexibility. Pauses appear mostly at standard syntactic boundaries rather than reflecting character deliberation, tactical hesitation, or dramatic weight.

Scoring prompt by sub-dimension (continued)

- **Holistic Integration:** The performance is recognizable and broadly coherent, but prosodic choices feel somewhat standard, detached, or inconsistent across scene transitions.

Score 4-3: Poor Prosodic Appropriateness
- **Pitch & Register:** Frequent mismatches between pitch contours and contextual intent. Baseline pitch register may feel unnatural for the persona, or intonation contours may conflict with the text (e.g., upward terminal inflections on definitive commands, or unmotivated pitch excursions).
- **Effort & Dynamics:** Vocal effort is poorly calibrated, sounding distinctly mismatched to the character's physical environment, distance from the listener, or emotional urgency (e.g., casual low effort during life-or-death confrontation, or continuous unvaried shouting).
- **Timing & Phrasing:** Speaking rate is erratic or rigidly mechanical. Phrasing frequently breaks across logical semantic units, and pauses appear random, disrupted, or artificially inserted, undermining character credibility.
- **Holistic Integration:** Prosodic cues repeatedly contradict the emotional and dramatic requirements of the scene, creating an artificial or disjointed impression.

Score 2-1: Severe Prosodic Inappropriateness
- **Pitch & Register:** Pervasive and severe disconnect between pitch contours and the dramatic context. The pitch contour may be totally flat and robotic during intense drama, wildly erratic and unmotivated, or entirely contrary to the target text and persona.
- **Effort & Dynamics:** Vocal effort completely ignores the communicative setting and emotional state, characterized by severe mismatch (e.g., deadpan delivery during extreme panic, or severe dynamic distortion unrelated to context).
- **Timing & Phrasing:** Timing, rhythm, and phrasing are entirely broken or inappropriate, displaying severe unnatural pacing, arbitrary halting pauses in the middle of core syntactic phrases, or an unyielding automated cadence.
- **Holistic Integration:** Total failure of prosodic appropriateness; the vocal delivery actively destroys the intended character, scene, and communicative meaning.

Notes:
1. Listen to the entire recording before deciding on a score.
2. Identify behavior that is actually audible and localize important evidence where possible.
3. Use the character profile, scene, and target text only to interpret the intended performance and any context-dependent criteria.
4. Apply the expression-mode-specific rubric and calibration consistently while respecting the dimension boundary.
5. Base the final score on the whole utterance and weigh isolated moments in proportion to their salience and coverage.
6. Actively check for score-limiting audible evidence instead of defaulting to a top-band score because the broad intent appears plausible or no gross mismatch is obvious. Do not invent limitations that are not audible.

For this quality axis, score_limiting_evidence means audible shortcomings, counterevidence, incomplete coverage, or local inconsistencies that constrain the score.

Before the score block, provide a concise English report containing:
- observed_behavior: the relevant behaviors heard for this dimension
- supporting_evidence: concrete acoustic and temporal evidence supporting the assigned level or quality
- score_limiting_evidence: concrete audible evidence limiting a higher score; write "No material score-limiting evidence observed" only when that is genuinely supported by the full recording
- rubric_application: how the evidence maps to the expression-mode-specific criteria
- score_rationale: why the selected score band and integer are justified

At the very end, output exactly this block with one integer from 1 to 10 and no extra text inside it:
<−−PROSODY_APPROPRIATENESS_SCORE_START−−>
prosody_appropriateness_score: [1-10]
<−−PROSODY_APPROPRIATENESS_SCORE_END−−>