# Qiuniu: Rubric-Based Expressive Understanding in Traditional Chinese Music

**Anonymous ACL submission**

## Abstract

Recent multimodal large language models (MLLMs) have achieved strong performance on audio-visual reasoning and music understanding tasks. However, existing benchmarks mainly evaluate observable signals such as objects, sounds, tempo, and loudness, while largely overlooking expressive meaning emerging from subtle multimodal patterns. We introduce **Qiuniu**, a benchmark for evaluating expressive understanding in Chinese traditional "Pipa" performance videos. Qiuniu contains expert annotations across five dimensions spanning technical correctness and expressive realization. Experiments on six recent MLLMs show that **all models remain far below expert-expert agreement**, with Gemini-2.5-Pro achieving the highest exact-match accuracy (0.411) yet still falling well short of the human ceiling (0.957). Repeatability experiments reveal that high score consistency often reflects score collapse rather than genuine evaluation, and expert verification confirms that model-cited evidence is frequently fabricated or ungrounded. These findings suggest that current MLLMs lack the stable, grounded musical assessment needed even at the technical level—a prerequisite for the higher-order expressive understanding that Qiuniu targets.

## 1 Introduction

Human interpreters of music rely on temporally coordinated multimodal signals, with expressive meaning emerging from cross-modal acoustic and visual structure rather than from isolated signals (Weck et al., 2024; Livingstone and Thompson, 2009a). Prior multimodal reasoning and affective communication studies similarly show that expressive meaning can emerge from implicit cross-modal cue integration without explicit semantic content (Zadeh et al., 2018; Lian et al., 2023b). However, existing multimodal benchmarks mainly evaluate observable signals and explicit semantic targets (Christop et al., 2026; Radevski et al., 2025; Li et al., 2025; Nguyen et al., 2025). In Traditional Chinese Music Scoring (TCMS), evaluation depends not only on technical correctness but also on whether a performance realizes the target expressive objective under the scoring rubric. Vocal- and lyric-free *pipa* performance provides a controlled setting in which models cannot rely on explicit verbal semantics, but must infer expressive intent from temporally coordinated multimodal cues spanning timbre, phrasing, dynamics, bodily motion, tempo, loudness, and articulation. It remains unclear whether **MLLMs can infer rubric-aligned expressive objectives from temporally coordinated multimodal cues under weak semantic supervision**.

Existing multimodal benchmarks mainly evaluate observable signals and explicit semantic targets, but do not jointly evaluate expressive meaning and technical–expressive separation. Audio-visual reasoning benchmarks such as DAVE (Radevski et al., 2025), OmniVideoBench (Li et al., 2025), AV-SpeakerBench (Nguyen et al., 2025), and ART (Christop et al., 2026) focus on temporal reasoning, audio-visual collaboration, speaker understanding, or audio reasoning from observable multimodal signals. Multimodal affect benchmarks including CMU-MOSEI (Zadeh et al., 2018) and MER2023 (Lian et al., 2023a) partially cover expressive meaning through sentiment or emotion recognition, but do not separate technical execution from expressive realization. Music benchmarks such as MuChoMusic (Weck et al., 2024), MusCaps (Manco et al., 2021), and SongEval (Yao et al., 2025) primarily evaluate perceived affect, caption generation, or aesthetic quality from salient acoustic patterns, often relying on lyrics, text supervision, or coarse acoustic correlates such as tempo. As summarized in Table 1, existing benchmarks therefore lack a controlled multimodal setting that jointly requires observable signal un-

| Benchmark | Domain | Modality | Task | Observable Signals | Expressive Meaning | Technical / Expressive Separation | Target |
|---|---|---|---|---|---|---|---|
| DAVE (Radevski et al., 2025) | Audio-visual reasoning | Video+Audio | QA | ✓ | ✗ | ✗ | Temporal reasoning |
| OmniVideoBench (Li et al., 2025) | Multimodal reasoning | Video+Audio | QA | ✓ | ✗ | ✗ | Audio-visual collaboration |
| AV-SpeakerBench (Nguyen et al., 2025) | Speech alignment | Video+Audio | QA | ✓ | ✗ | ✗ | Speaker understanding |
| ART (Christop et al., 2026) | Audio reasoning | Audio | QA | ✓ | ✗ | ✗ | Audio reasoning |
| CMU-MOSEI (Zadeh et al., 2018) | Multimodal affect | V+A+T | Sentiment | ✓ | △ | ✗ | Sentiment prediction |
| MER2023 (Lian et al., 2023a) | Multimodal affect | V+A+T | Emotion | ✓ | △ | ✗ | Emotion recognition |
| MuChoMusic (Weck et al., 2024) | Music | Audio+Text | QA / Cls | ✓ | △ | ✗ | Emotion / attribute |
| MusCaps (Manco et al., 2021) | Music | Audio | Captioning | ✓ | ✗ | ✗ | Music caption |
| SongEval (Yao et al., 2025) | Music (generated) | Audio | Aesthetic rating | ✓ | △ | ✗ | Song aesthetics |
| **Qiuniu (Ours)** | Music performance | Video+Audio | Rubric eval | ✓ | ✓ | ✓ | **Expressive understanding** |

Table 1: Comparison of existing multimodal benchmarks across evaluated capabilities. The table contrasts prior benchmarks and Qiuniu in terms of observable signal evaluation, expressive meaning, and technical–expressive separation. Symbols denote the level of coverage: ✓ indicates explicit evaluation, ✗ indicates no evaluation, and △ denotes partial or indirect coverage.

derstanding, expressive meaning inference, and technical–expressive separation under weak semantic supervision.

To address this gap, we propose Qiuniu, a benchmark grounded in TCMS for evaluating expressive understanding in instrumental audio–visual *pipa* performance. Qiuniu contains 115 performance videos spanning 42 canonical pieces and 539.6 minutes of video, where each sample is paired with annotated sheet music specifying target emotion and imagery. The dataset includes both expressive and deliberately restrained performances to reduce reliance on salient acoustic or visual cues. Following standardized TCMS grading criteria used in examinations and performance assessment, we define a five-dimensional ordinal rubric covering Pitch & Rhythm, Technical Skill, Completeness, Musical Flow, and Expressivity. We collect expert annotations on a 1–3 scale for each dimension and report inter-annotator agreement (macro-average: Acc. 0.957, $\kappa$ 0.925, AC2 0.958). This setup requires models to infer expressive realization from temporally coordinated multimodal cues rather than isolated acoustic or visual patterns. Based on this setup, we formulate the task as rubric-based performance grading, where a model predicts ordinal rubric scores from audio–visual performance and annotated sheet music. Evaluation uses exact-match accuracy and within-one-level agreement ($|\Delta| \leq 1$), reported per dimension and macro-averaged across dimensions.

Our evaluation shows that current MLLMs do not understand musical emotion as an adequacy judgment. In Qiuniu, the question is not whether a performance sounds expressive in general, but whether it realizes the specific emotion, imagery, and artistic conception required by the score. Across six representative MLLMs, all models remain far below expert-expert agreement. Gemini-2.5-Pro performs best with an exact-match accuracy of 0.411, still far below the human ceiling of 0.957. This gap cannot be explained by a simple technical versus expressive split, because in TCMS, technique is part of how emotion is realized through tone, phrasing, timing, breath, and movement. Instead, models often replace expressive adequacy with generic proxies such as fluency, tempo stability, visible engagement, or positive evaluative language. Repeatability and expert verification further show that stable outputs often reflect score collapse, and that cited evidence is frequently fabricated or ungrounded. These findings suggest that current MLLMs can imitate the language of expert evaluation, but they cannot reliably judge whether audio-visual performance cues fulfill a score-specified emotional objective. Our contributions are summarized as follows:

- We propose **Qiuniu**, the first audio–visual benchmark grounded in Traditional Chinese Music Scoring (TCMS) for evaluating expressive understanding in instrumental *pipa* performance. Qiuniu contains 115 performance videos spanning 42 canonical pieces with expert-annotated five-dimensional ordinal scores (macro-average $\kappa$ 0.925), providing a controlled setting for expressive objective inference under weak semantic supervision.

- We define **rubric-based performance grading**, a multimodal evaluation task in which models predict five-dimensional ordinal TCMS scores from audio–visual performance and annotated sheet music. The task requires joint assessment of technical execution and expressive realization from temporally coordinated multimodal cues, enforcing technical–expressive separation absent from existing benchmarks.

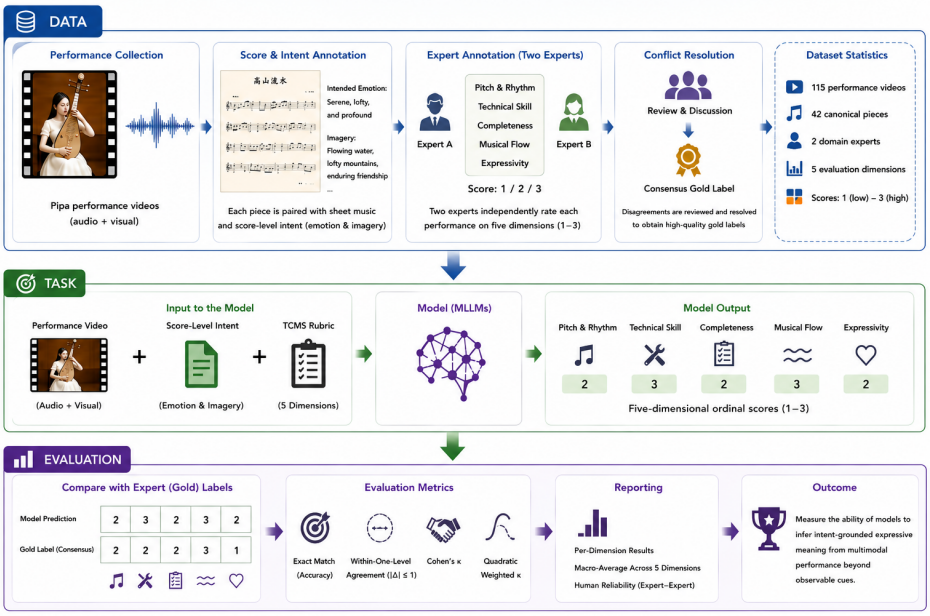


Figure 1: Overview of the Qiuniu benchmark. Using vocal- and lyric-free *pipa* performance video, score-level intent annotation, and the TCMS rubric, MLLMs predict five-dimensional ordinal ratings covering both technical and expressive performance quality.

- We benchmark six representative MLLMs and find that all models remain substantially below expert-level agreement (best: Gemini-2.5-Pro at 0.411 vs. human ceiling of 0.957). Models frequently rely on shallow multimodal correlates such as fluency, tempo, or visible engagement, while producing unstable or ungrounded expressive judgments.

## 2 Related Work

**Expressive Understanding and Multimodal Reasoning.** Human expressive understanding emerges from coordinated multimodal structure rather than isolated observable cues (Koffka, 1935; Wertheimer, 1923; Livingstone and Thompson, 2009a). Prior work further suggests that affective and abstract meaning can be inferred indirectly from multimodal sensory patterns without explicit semantics (Lakoff and Johnson, 1980). Existing multimodal affective benchmarks, including CMU-MOSEI and MER (Zadeh et al., 2018; Lian et al., 2023a,b), primarily evaluate directly observable emotions from speech, text, facial expression, and audio. Recent MLLMs such as GPT-4o, Gemini, and SALMONN (OpenAI et al., 2024; Team et al., 2025; Tang et al., 2024a) demonstrate increasingly strong multimodal reasoning across visual, audio, and temporal domains. Corresponding benchmarks including DAVE (Radevski et al., 2025), OmniVideoBench (Li et al., 2025), AV-SpeakerBench (Nguyen et al., 2025), and ART (Christop et al., 2026) evaluate modality dependence, audio–visual collaboration, speaker-aware alignment, and audio reasoning, but mainly focus on observable signals, explicit semantic targets, or temporal-causal reasoning.

**Music Understanding and Performance Evaluation.** Existing music benchmarks mainly evaluate captioning, affect recognition, or aesthetic quality from salient acoustic correlates. MusCaps studies music caption generation (Manco et al., 2021), MuChoMusic evaluates music understanding across mood and expression while revealing substantial reliance on language shortcuts (Weck et al., 2024), and SongEval introduces multidimensional rubric-based evaluation for generated song aesthetics (Yao et al., 2025). Existing musician assessment systems further emphasize technical correctness and observable execution quality (Evin, 2024). Across these settings, expressive understanding is largely operationalized through observable correlates such as tempo, loudness, fluency, or perceived affect, rather than score-conditioned expressive inference under weak semantic supervision.

## 3 Qiuniu

We introduce **Qiuniu**, a benchmark grounded in Traditional Chinese Music Scoring (TCMS). As illustrated in Figure 1, Qiuniu jointly evaluates technical correctness and expressive realization through a five-dimensional expert rubric, enabling

controlled analysis of whether current multimodal systems can align subtle sound and performance patterns with intended expressive meaning.

### 3.1 Task Definition

We define Qiuniu as a multimodal ordinal grading task for expressive music performance evaluation. Given an audio–visual *pipa* performance sample $x = (v, a)$, where $v$ denotes the video stream and $a$ denotes the synchronized audio stream, the model predicts rubric-based scores over five evaluation dimensions:

$$\mathcal{D} = \{d_{\text{pitch}}, d_{\text{tech}}, d_{\text{complete}}, d_{\text{flow}}, d_{\text{expr}}\},$$

corresponding to *Pitch & Rhythm*, *Technical Skill*, *Completeness*, *Musical Flow*, and *Expressivity*.

For each dimension $d_i \in \mathcal{D}$, the model predicts an ordinal score

$$y_i \in \{1, 2, 3\},$$

where higher scores indicate stronger alignment with the TCMS evaluation rubric. The final prediction is therefore a five-dimensional ordinal vector

$$\mathbf{y} = (y_1, y_2, y_3, y_4, y_5).$$

The task requires models to infer rubric-aligned expressive objectives from temporally coordinated multimodal performance signals, rather than relying on isolated acoustic or visual cues.

### 3.2 Dataset Curation and Annotation

**Data Source and Collection.** Qiuniu was constructed through domain-expert annotation rather than crowdsourcing.[1] The dataset consists of vocal- and lyric-free *pipa* performance videos drawn from graded examination repertoire spanning Levels 1–11. In total, Qiuniu contains 115 performance videos covering 42 canonical pieces. Each video is paired with the corresponding sheet music and score-level intent annotation specifying the intended emotion and imagery of the piece. The collection covers students with varying proficiency levels and includes both emotionally expressive and deliberately restrained performances, including cases where technical execution is relatively mature but expressive communication remains limited. This design reduces reliance on trivial shortcuts such as assigning uniformly low scores to weak performances or using salient cues such as loudness, tempo, or exaggerated body movement. Full dataset statistics are provided in Appendix A.1.

[1] All participants involved in data collection and annotation signed informed consent agreements.

**Annotation Protocol.** We recruited two domain experts with professional *pipa* training, more than ten years of teaching experience, and prior experience as evaluators in graded examinations or comparable professional assessment settings. The annotation rubric was co-designed by computer science researchers and Traditional Chinese Music educators, and covers five TCMS dimensions: *Pitch & Rhythm*, *Technical Skill*, *Completeness*, *Musical Flow*, and *Expressivity*. Detailed scoring guidelines, examples, and boundary cases are provided in Appendix B.2. During annotation, each expert independently rated every sample on a 1–3 ordinal scale for each dimension, using the score-level intent annotation and rubric as reference. The two experts completed annotation separately before any reconciliation, preventing direct influence during the first-pass labeling stage.

**Conflict Resolution and Gold Labels.** After independent annotation, we identified all cases where the two experts assigned different scores on any dimension. Disagreements were first checked for procedural issues, including missing video segments, unclear audio quality, or ambiguous alignment with the intended piece; samples with unresolved data-quality problems were excluded from the finalized benchmark. For valid samples, experts performed a second independent review focused only on the disagreed dimensions, using the original rubric and score-level intent annotation. Remaining disagreements were resolved through a moderated discussion organized by the research team, where experts were asked to justify their scores with reference to observable performance evidence and TCMS criteria. The final gold label was assigned only after consensus was reached; otherwise, the case was excluded from the gold-label set. This process ensures that reference labels reflect stable expert judgment rather than a single annotator's preference.

**Annotation Reliability.** Table 2 reports inter-annotator agreement before consensus resolution. The two experts show high agreement across all five dimensions, with a macro-average Accuracy of 0.957, Cohen's $\kappa$ of 0.925, and AC2 of 0.958. These results indicate that the rubric supports consistent expert judgment and that the reconciled labels are suitable as reference labels for model evaluation.

Table 2: Inter-annotator agreement across the five TCMS dimensions before consensus resolution.

| Dimension | Acc. | $\kappa$ | AC2 |
|---|---|---|---|
| Pitch & Rhythm | 0.939 | 0.895 | 0.941 |
| Technical Skill | 0.957 | 0.925 | 0.958 |
| Completeness | 0.930 | 0.879 | 0.933 |
| Musical Flow | 1.000 | 1.000 | 1.000 |
| Expressivity | 0.957 | 0.927 | 0.957 |
| **Macro Avg.** | **0.957** | **0.925** | **0.958** |

### 3.3 Evaluation

We evaluate a diverse set of representative MLLMs spanning closed-source API systems and open-source multimodal models, including general-purpose vision–language systems and omni-modal audio–visual models (Gemini Team, Google, 2025; Xu et al., 2025; Tang et al., 2024b; OpenBMB Team, 2025; Abouelenin et al., 2025; Xiaomi MiMo Team, 2026). The full model list is provided in Table 3.

| Model | Family | Params | Modality | Access |
|---|---|---|---|---|
| Gemini-2.5-Pro | Gemini | – | Audio + Video | API |
| Qwen-Omni | Qwen | 7B | Audio + Video | Open-source |
| SALMONN-7B | SALMONN | 7B | Audio + Video | Open-source |
| MiMo-V2-Omni | MiMo | 7B | Audio + Video | API / Publicly available |
| MiniCPM-o-4.5 | MiniCPM | 8B | Audio + Video | Open-source |
| Phi-4-Multimodal | Microsoft Phi | 5.6B | Audio + Video | Open-source |

Table 3: Audio-visual multimodal models evaluated in our experiments.

All models are evaluated under a unified zero-shot protocol following the task formulation in Section 3.1. For each sample, models receive the audio–visual performance, score-level intent annotation, and TCMS rubric, and predict a five-dimensional ordinal rating vector. All models use the same prompt template, scoring instructions, and constrained output format described in Appendix C.1. Closed-source models are accessed through public APIs under default settings, while open-source models use official inference frameworks.

We compare model predictions against expert reference ratings averaged across the two annotators unless otherwise specified. Since the task is a multi-dimensional ordinal grading problem, we use quadratic-weighted Cohen's $\kappa$ as the primary metric. We additionally report exact-match Accuracy and within-one-level agreement ($|\Delta| \leq 1$) as complementary measures of strict and near-miss consistency. All metrics are reported per dimension and macro-averaged across the five rubric dimensions. Expert–expert agreement is reported separately as a human reliability reference.

## 4 Experiments

### 4.1 Main Results

Table 4: Full per-dimension agreement between expert ratings and model predictions. PR: Pitch & Rhythm; TC: Tone Control; CAT: Completeness / Accuracy / Technical Proficiency; MFB: Musical Flow & Breath Control; ME: Musical Expressiveness.

| Model | Dim. | Exact | Within-1 | $\kappa$ | Quad. $\kappa$ |
|---|---|---|---|---|---|
| Gemini-2.5-Pro | PR | 0.383 | 0.887 | −0.042 | −0.129 |
| | TC | 0.435 | 0.896 | 0.027 | −0.080 |
| | CAT | 0.417 | 0.896 | 0.017 | −0.071 |
| | MFB | 0.409 | 0.896 | −0.024 | −0.121 |
| | ME | 0.409 | 0.896 | 0.000 | −0.040 |
| MiniCPM-o-4.5 | PR | 0.426 | 0.948 | 0.045 | −0.038 |
| | TC | 0.400 | 0.983 | −0.029 | 0.024 |
| | CAT | 0.357 | 0.983 | −0.059 | 0.021 |
| | MFB | 0.409 | 0.983 | −0.009 | 0.033 |
| | ME | 0.348 | 0.991 | −0.037 | −0.026 |
| MiMo-V2-Omni | PR | 0.369 | 0.937 | 0.002 | 0.004 |
| | TC | 0.387 | 0.937 | −0.026 | −0.079 |
| | CAT | 0.279 | 0.964 | −0.096 | 0.050 |
| | MFB | 0.369 | 0.928 | −0.016 | −0.070 |
| | ME | 0.423 | 0.937 | 0.082 | 0.061 |
| Qwen-Omni | PR | 0.306 | 0.917 | 0.003 | −0.020 |
| | TC | 0.352 | 0.917 | 0.021 | −0.019 |
| | CAT | 0.269 | 0.917 | −0.027 | −0.035 |
| | MFB | 0.306 | 0.907 | −0.005 | −0.054 |
| | ME | 0.324 | 0.926 | −0.027 | −0.037 |
| SALMONN-7B | PR | 0.283 | 0.867 | 0.003 | 0.003 |
| | TC | 0.124 | 0.478 | 0.013 | −0.019 |
| | CAT | 0.150 | 0.451 | 0.028 | −0.005 |
| | MFB | 0.133 | 0.558 | −0.004 | 0.023 |
| | ME | 0.186 | 0.841 | −0.120 | 0.071 |
| Phi-4 | PR | 0.113 | 0.443 | 0.000 | 0.000 |
| | TC | 0.096 | 0.461 | 0.000 | 0.000 |
| | CAT | 0.113 | 0.426 | 0.000 | 0.000 |
| | MFB | 0.087 | 0.435 | 0.000 | 0.000 |
| | ME | 0.122 | 0.478 | 0.000 | 0.000 |

Table 5: Main results on Qiuniu. We report exact-match accuracy, within-one-level agreement ($|\Delta| \leq 1$), Cohen's $\kappa$, and quadratic weighted $\kappa$ between model predictions and expert ratings across five TCMS dimensions. Dimension abbreviations: PR = Pitch & Rhythm; TC = Tone Control; CAT = Completeness / Accuracy / Technical Proficiency; MFB = Musical Flow & Breath Control; ME = Musical Expressiveness.

Table 6 presents macro-average results across all models, ranked by exact-match accuracy. Tech and Expr denote average Exact Accuracy over the four technical dimensions and Musical Expressivity, respectively; the human row provides the expert–expert reliability ceiling. Full per-dimension breakdowns are provided in Appendix D (Table 4).

Across all evaluated systems, every model remains substantially below expert-expert agreement (Exact: 0.957, $\kappa$: 0.925). Gemini-2.5-Pro achieves the highest exact-match accuracy (0.411), followed by MiniCPM-o-4.5 (0.388) and MiMo-V2-Omni (0.365). SALMONN-7B and Phi-4 perform sub-

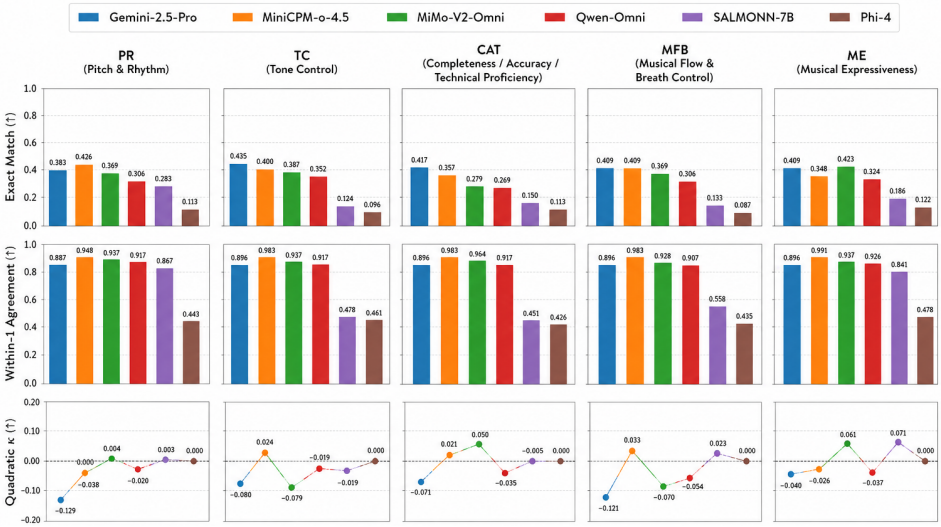


Figure 2: Model–expert agreement across five TCMS dimensions on Qiuniu. Higher is better for all metrics.

Table 6: Macro-average model–expert agreement across five dimensions.

| Model | Rank | Exact | W-1 | $\kappa$ | Tech | Expr |
|---|---|---|---|---|---|---|
| Human | – | 0.957 | – | 0.925 | – | – |
| Gemini-2.5-Pro | 1 | 0.411 | 0.894 | −0.004 | 0.411 | 0.409 |
| MiniCPM-o-4.5 | 2 | 0.388 | 0.978 | −0.018 | 0.398 | 0.348 |
| MiMo-V2-Omni | 3 | 0.365 | 0.941 | −0.011 | 0.351 | 0.423 |
| Qwen-Omni | 4 | 0.311 | 0.917 | −0.007 | 0.308 | 0.324 |
| SALMONN-7B | 5 | 0.175 | 0.639 | −0.016 | 0.172 | 0.186 |
| Phi-4 | 6 | 0.106 | 0.449 | 0.000 | 0.102 | 0.122 |

stantially lower, with SALMONN-7B showing a strikingly low within-one agreement of 0.639 compared to above 0.894 for all stronger models, indicating that its predictions frequently fall outside even a one-point margin of the expert reference. All $\kappa$ values are near zero or negative, confirming that no model achieves meaningful agreement beyond chance under the ordinal evaluation protocol.

The ranking itself, however, must be interpreted with caution. As detailed in Section 4.3, the top-ranking models by Exact Accuracy frequently achieve their scores through score collapse—assigning a narrow range of scores (e.g., 2/2/2/2/2 or 3/3/3/3/3) uniformly across samples rather than making genuinely calibrated judgments. This means that higher accuracy reflects a fortuitous match between a model's default output range and the expert score distribution, rather than reliable expert-aligned evaluation.

Examining the Tech and Expr columns reveals a mixed picture across model families. MiniCPM-o-4.5 shows higher accuracy on technical dimensions than on *Musical Expressivity* (0.398 vs. 0.348), consistent with a technical–expressive gap. For Gemini-2.5-Pro, the two are nearly equal (0.411 vs. 0.409). By contrast, MiMo-V2-Omni and other models show the opposite pattern, with higher expressivity accuracy than technical average, a finding we discuss further in Section 4.2. Taken together, the results suggest that the technical–expressive gap is not universal across models, and that the primary bottleneck may differ depending on a model's default scoring behavior and calibration failure mode.

### 4.2 Technical Competence vs. Expressive Intent

Table 6 includes a direct comparison of average Exact Accuracy on the four technical dimensions (Tech) versus *Musical Expressivity* (Expr). The picture is model-dependent rather than uniform. MiniCPM-o-4.5 shows a clear technical advantage (Tech: 0.398 vs. Expr: 0.348), and Gemini-2.5-Pro shows near-equal performance (0.411 vs. 0.409). By contrast, MiMo-V2-Omni scores higher on expressivity than on technical dimensions (0.351 vs. 0.423), and the pattern for SALMONN-7B and Phi-4 is similarly mixed.

This heterogeneity is partly explained by score collapse behavior. Models that systematically assign high scores (e.g., Qwen-Omni, MiniCPM) happen to align better with expert technical scores because expert technical scores are on average higher than expressivity scores in our dataset, producing an apparent technical advantage that reflects distributional alignment rather than genuine capability. Conversely, models with other collapse patterns

may appear stronger on expressivity for the same artifactual reason. In both cases, the high within-one agreement (above 0.89 for the top four models) combined with near-zero $\kappa$ confirms that what looks like dimensional accuracy is largely driven by range matching, not reliable scoring.

What is consistent across all models is that none achieves meaningful agreement with expert raters on *any* dimension. The repeatability and case study evidence in Sections 4.3 and E.2 further shows that model-generated justifications—including descriptions of technical aspects—are frequently fabricated or grounded in non-musical events. This means the apparent dimensional accuracy differences cannot be reliably attributed to genuine technical versus expressive perception gaps. Rather, the dominant finding is that current MLLMs lack stable, grounded evaluation capability across all five dimensions, and that the path toward reliable expressive understanding requires solving the more fundamental problem of grounded musical assessment first.

### 4.3 Repeatability and Evidence Reliability Analysis

To further diagnose the sources of model failure on Qiuniu, we conduct a repeatability experiment grounded in our staged evaluation design. For each evaluated model, we run the same evaluation prompt on the same performance video three times independently, recording scores, cited time segments, and supporting evidence for each run. We measure three types of consistency across runs: (1) *score consistency*, the proportion of samples for which the five-dimensional scores are identical across all three runs; (2) *timestamp consistency*, measured both as exact match and as segment overlap, capturing whether the time intervals cited as evidence are stable; and (3) *evidence consistency*, whether the described technical issues and reasoning are substantively the same across runs. Table 7 summarizes the quantitative consistency metrics. Detailed failure patterns and expert verification status are provided in Appendix E.1.

As shown in Table 7, models fall into two groups: several models (SALMONN-7B, Qwen-Omni, Phi-4, MiniCPM-o-4.5) achieve high consistency across all three metrics, while others (MiMo-V2-Omni, Gemini-2.5-Pro) are unstable across runs. Crucially, high consistency is not evidence of reliable evaluation. To understand what the stable outputs are actually based on, we analyze the rationales generated by the high-consistency models. We find that these models do not ground their scores in rubric-specified expressive objectives; instead, they repeatedly invoke a small set of generic proxy signals—surface fluency, tempo stability, visible performer engagement, and positive evaluative language—and reuse these same justifications across structurally different performances. In other words, the stability arises because the model applies a fixed, content-independent scoring template rather than because it consistently perceives the same musical content.

Table 7: Repeatability metrics across three runs.

| Model | Score | Full | T-Exact | T-Over. | Evid. |
|---|---|---|---|---|---|
| SALMONN-7B | 100 | 100 | 99.1 | 99.5 | 98.6 |
| Qwen-Omni | 99.1 | 98.2 | 37.1 | 92.6 | 97.1 |
| MiMo-V2-Omni | 76.0 | 53.2 | 0.0 | 6.7 | 37.5 |
| Gemini-2.5-Pro | 36.0 | 30.4 | 0.0 | 4.9 | 0.0 |
| Phi-4 | 100 | 100 | 83.5 | 83.5 | 100 |
| MiniCPM-o-4.5 | 100 | 100 | 100 | 100 | 100 |

We then ask domain experts to verify the authenticity of the evidence cited by these stable models. The verification reveals a consistent pattern: when a model's output is stable, its interpretation typically deviates from the score-specified intent, and the supporting evidence it cites is frequently fabricated or unsupported by the actual performance. Experts report that cited time segments often do not contain the claimed technical issues, that technical terminology is misapplied, and in several cases that the cited events did not occur at all (detailed model rationales and expert verdicts are provided as case studies in Appendix E.2). Stable repeatability therefore does not indicate grounded evaluation; rather, it indicates that the model has settled on a confident but ungrounded scoring heuristic that bypasses genuine assessment of whether the performance fulfills the rubric objective.

These observations motivate a deeper re-examination of the overall results. Before attributing model failure on *Musical Expressivity* to a deficit in Theory of Mind-like inference, it is necessary to consider whether current models possess the prerequisite capabilities for musical evaluation in the first place. The repeatability analysis reveals that current MLLMs do not function as stable musical evaluators even on technical dimensions: score collapse, timestamp hallucination, and run-to-run instability are widespread. This raises the question of whether the observed technical–expressive gap reflects a specific failure of expressive understand-

ing, or whether it partially reflects a more general failure of grounded, calibrated musical assessment at the perceptual level.

However, we note that technical evaluation failure and expressive evaluation failure need not be fully coupled. Research on congenital amusia is informative in this regard: individuals with amusia, who show impaired processing of pitch and musical structure at a perceptual level, nonetheless demonstrate intact or near-intact emotion recognition from music in certain paradigms (Tillmann et al., 2023; Hoarau et al., 2024). More directly, Lévêque et al. (2018) demonstrate that amusic individuals can rate the perceived intensity of musical emotions as accurately as controls, even when they fail to identify the emotion category. Under this dissociation, impaired technical perception does not necessarily entail impaired emotion perception. This suggests that even if current MLLMs fail to evaluate technical dimensions reliably, their failure on *Musical Expressivity* may reflect additional deficits beyond musical perception alone, including the absence of intent-grounded inference, cultural associative reasoning, or the multimodal empathic mechanisms described by Livingstone and Thompson (2009b). Specifically, Livingstone and Thompson (2009b) argue that musical emotion perception engages not only acoustic processing but also empathic resonance, mirror-system activation, and culturally learned associations between performance patterns and emotional meaning.

Taken together, these results support a two-level interpretation of model failure on Qiuniu. At the first level, current MLLMs lack stable and grounded technical evaluation capabilities, evidenced by score collapse, timestamp hallucination, and output instability. At the second level, even if technical evaluation were more reliable, the expressive gap would likely persist, because mapping multimodal performance behavior to intended expressive meaning requires higher-order inference that extends beyond recognizing technical cues. The combination of these two failure modes suggests that improving performance on Qiuniu will require advances not only in musical perception and calibrated evaluation, but also in the kind of intent-grounded and culturally situated reasoning that human expert evaluators bring to musical assessment.Qualitative case studies further reveal that models frequently generate musically plausible but weakly grounded rationales, including unsupported references to performance techniques and expressive intentions.

## 5 Conclusion and Future Work

In this work, we introduce **Qiuniu**, a multimodal benchmark for evaluating expressive understanding in music performance. Unlike existing multimodal benchmarks that mainly focus on perception-oriented capabilities such as audio-visual recognition or explicit semantic understanding, Qiuniu explicitly separates technical correctness from expressive realization within the same evaluation framework. Experiments across six MLLMs show that all models remain far below expert-expert agreement, with Gemini-2.5-Pro achieving the best overall exact-match accuracy (0.411) but still well short of the human ceiling (0.957). Contrary to initial expectations, the technical–expressive gap is not uniform across models: apparent dimensional differences are largely explained by score collapse behavior rather than genuine capability differences. Repeatability experiments further reveal that high score consistency frequently reflects degenerate scoring heuristics, and expert verification confirms that model-cited timestamps and technical evidence are often fabricated.

Notably, this failure at the technical evaluation level does not straightforwardly explain the expressive gap. Research on congenital amusia suggests that technical musical perception and emotional evaluation can be dissociated: individuals with impaired pitch processing can still match emotional intensity in music as accurately as controls (Lévêque et al., 2018). This implies that even if current MLLMs were better grounded in technical perception, the expressive gap might persist, because evaluating intended expressive meaning requires additional capacities—including intent-grounded inference, cultural associative reasoning, and the multimodal empathic mechanisms described by Livingstone and Thompson (2009b)—that go beyond observable signal recognition.

Future work may benefit from hierarchical temporal modeling, intent-aware multimodal alignment, and music-centered multimodal pretraining that targets grounded evaluation rather than description generation. More broadly, we hope Qiuniu motivates future benchmarks that evaluate not only perceptual recognition but also the higher-order, culturally situated understanding that expert musical assessment requires.

## Ethical Considerations

Qiuniu contains audio-visual recordings of human musical performances and therefore involves both human participation and culturally grounded artistic evaluation. All performance videos were collected with informed consent from the performers or their legal guardians when appropriate, and participants were informed that the recordings could be used for academic research and benchmark development. To reduce privacy risks, all released samples are anonymized and exclude personally identifiable information such as names, affiliations, and contact details. The benchmark is constructed strictly for non-commercial academic research purposes. The musical repertoire is derived from publicly distributed instructional and graded examination materials widely used in Chinese music education, and no copyrighted instructional material is redistributed beyond what is necessary for evaluation and educational reference. Future releases of the dataset should continue to comply with applicable copyright regulations and educational fair-use policies.

A more fundamental ethical consideration concerns the inherently interpretive nature of expressive evaluation. Unlike tasks grounded in directly observable labels, musical expressiveness depends heavily on cultural conventions, pedagogical traditions, and human artistic judgment. Although Qiuniu adopts a standardized TCMS rubric and uses experienced expert annotators with high inter-rater agreement, expressive meaning cannot be fully reduced to objective measurement. As a result, the benchmark should not be interpreted as defining a universal notion of musical quality or emotional correctness, but rather as reflecting a specific expert evaluation framework grounded in Traditional Chinese Music pedagogy. More broadly, our findings should not be interpreted as evidence that current MLLMs are incapable of artistic or cultural understanding in general. Instead, Qiuniu highlights a specific limitation in consistently aligning subtle multimodal performance patterns with intended expressive meaning. We therefore discourage the direct deployment of current systems for high-stakes artistic assessment, educational grading, or professional certification without substantial human oversight.

## Limitations

While Qiuniu provides a new benchmark for studying expressive understanding in multimodal systems, several limitations remain. First, the current benchmark focuses exclusively on *pipa* performance within the TCMS framework. Although this design enables controlled analysis of expressive evaluation under a well-defined pedagogical setting, it also limits the scope of cultural and musical generalization. Expressive conventions differ substantially across instruments, musical traditions, and performance cultures. In addition, the dataset remains relatively small compared with large-scale multimodal benchmarks, containing 115 performance videos covering 42 canonical pieces annotated by two expert evaluators. Although expert agreement is consistently high, the limited dataset scale constrains statistical power and restricts the diversity of expressive styles, performer backgrounds, and interpretive variation represented in the benchmark.

Another limitation concerns the formulation of expressiveness itself. In Qiuniu, expressive understanding is operationalized through ordinal ratings under a standardized rubric. While this improves annotation consistency and evaluation comparability, expressive meaning in music is often continuous, context-dependent, and temporally evolving. Certain performances may reasonably admit multiple valid interpretations even among experts, meaning that benchmark scores should be interpreted as expert-aligned judgments within a specific evaluative framework rather than absolute ground truth. Finally, the current study evaluates only a limited set of recent multimodal systems under a unified prompting protocol. The observed performance gap may partially depend on prompting strategy, audio preprocessing, inference stochasticity, or model-specific alignment procedures. Future work should therefore investigate whether specialized music-centered multimodal pretraining, temporal reasoning mechanisms, or intent-aware alignment objectives can improve expressive understanding beyond observable signal recognition.

## References

Amr Abouelenin and 1 others. 2025. Phi-4-multimodal technical report. *Preprint*, arXiv:2503.01743.

Iwona Christop, Mateusz Czyżnikiewicz, Paweł Skórzewski, Łukasz Bondaruk, Jakub Kubiak,

Marcin Lewandowski, and Marek Kubis. 2026. A benchmark for audio reasoning capabilities of multimodal large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*.

Munni Evin. 2024. A review on AI-enabled techniques for evaluating musician's performance. In *AIP Conference Proceedings*, volume 3149, page 140018.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Charlotte Hoarau, Agathe Pralus, Annie Moulin, Nathalie Bedoin, Julien Ginzburg, Lesly Fornoni, Pierre-Emmanuel Aguera, Barbara Tillmann, and Anne Caclin. 2024. Deficits in congenital amusia: Pitch, music, speech, and beyond. *Neuropsychologia*, 202:108960.

Kurt Koffka. 1935. *Principles of Gestalt Psychology*. Harcourt, Brace and Company, New York.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Yohana Lévêque, Perrine Teyssier, Patrick Bouchet, Emmanuel Bigand, Anne Caclin, and Barbara Tillmann. 2018. Musical emotions in congenital amusia: Impaired recognition, but preserved emotional intensity. *Neuropsychology*, 32(7):880–894.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, Ying He, and 1 others. 2025. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*.

Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023a. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10):1440.

Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023b. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10):1440.

Steven Livingstone and William Thompson. 2009a. The emergence of music from the theory of mind. *Musicae Scientiae*, Special Issue 2009/10:83–115.

Steven Robert Livingstone and William Forde Thompson. 2009b. The emergence of music from the theory of mind. *Musicae Scientiae*, pages 83–115.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2021. Muscaps: Generating captions for music audio. *arXiv*, abs/2104.11984.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and Yong Jae Lee. 2025. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenBMB Team. 2025. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming. Technical blog and model release. Accessed May 24, 2026.

Gorjan Radevski, Teodora Popordanoska, Matthew B. Blaschko, and Tinne Tuytelaars. 2025. Dave: Diagnostic benchmark for audio visual evaluation. *arXiv preprint arXiv:2503.09321*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024a. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024b. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Barbara Tillmann, Julia E. Graves, Francesca Talamini, Yohana Lévêque, Lesly Fornoni, Charlotte Hoarau, Agathe Pralus, Julien Ginzburg, Philippe Albouy, and Anne Caclin. 2023. Auditory cortex and beyond: Deficits in congenital amusia. *Hearing Research*, 437:108855.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio–language models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, San Francisco, USA.

Max Wertheimer. 1923. Untersuchungen zur lehre von der gestalt ii. *Psychologische Forschung*, 4:301–350.

Xiaomi MiMo Team. 2026. Mimo-v2-omni. Model card. Accessed May 24, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, Hao Liu, and Lei Xie. 2025. SongEval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*.

Amir Zadeh, Paul Pu Liang, Jonathan Van Briesen, Soujanya Poria, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2236–2246. Association for Computational Linguistics.

# A Dataset

## A.1 Dataset Statistics

Table 8 provides complete dataset statistics, including the distribution of performances by technical and expressive proficiency and key scale metrics.

## A.2 Dataset Details

The dataset consists of **115 pipa performance videos** collected from student practice and assessment settings. All performers are students with formal training backgrounds. To preserve privacy, no personally identifying information is included. The repertoire is drawn from standardized instructional and examination materials widely used in Chinese music education, including:

- *Juvenile Pipa Course* (Shanghai Music Publishing House)
- *Pipa Grading Examination Materials* (Hubei Musicians Association)
- *Pipa Scores Collection* (People's Music Publishing House)
- *National Pipa Examination Repertoire* (China Federation of Literary and Art Circles Press)

These pieces are commonly used in formal grading examinations, ensuring that the performances reflect standardized evaluation criteria and pedagogical objectives in Chinese traditional music training.

## A.3 Performer Information

To protect privacy, all performers in the Qiuniu dataset are anonymized. Each performer is represented by an anonymous identifier. The dataset includes student performers from multiple musical training institutions in China, spanning a wide range of proficiency levels corresponding to graded examination repertoires (Levels 1–10). For each performance sample, the metadata records:

- Anonymous performer ID
- Musical piece
- Examination level
- Expert evaluation scores across five dimensions

No personally identifiable information (e.g., real names, institutions, or demographic details) is included in the released dataset.

# B Annotation Setup

## B.1 TCMS Scoring Rubric

The TCMS evaluation rubric defines five dimensions for assessing pipa performance, each scored on a 3-point ordinal scale (3 = Excellent, 2 = Acceptable, 1 = Poor).

**(1) Pitch & Rhythm**

- 3: Accurate pitch and stable rhythm, including correct execution of complex rhythmic patterns
- 2: Generally correct pitch with noticeable deviations; rhythm framework preserved but unstable
- 1: Frequent pitch errors and unstable rhythm with repeated timing issues

**(2) Tone Control**

- 3: Clean and controlled tone with characteristic timbre of traditional music
- 2: Acceptable tone with noticeable noise or dryness
- 1: Poor tone quality with significant noise or weak timbre

**(3) Completeness, Accuracy & Technique**

- 3: Fluent and complete performance with accurate execution of techniques

| Performance Distribution | Definition | # Videos |
|---|---|---|
| Technical strong, Emotion weak | Tech $\geq$2.5 & Emotion $\leq$2 | 24 |
| Emotion strong, Technical weak | Emotion $\geq$2.5 & Tech $\leq$2 | 9 |
| Both strong | Tech $\geq$2.5 & Emotion $\geq$2.5 | 46 |
| Both weak | Tech $\leq$2 & Emotion $\leq$2 | 23 |
| Other / transitional cases | Not covered by the above categories | 13 |
| Total videos | | 115 |
| Unique musical pieces | | 42 |
| Total duration | | 539.6 min ($\approx$9.0 hours) |
| Average duration per sample | | 284.0 s ($\approx$4.7 min) |
| Difficulty levels | | 1–10 |
| Annotation dimensions | | 5 |
| Score scale | | 1–3 |
| Annotators | | 2 experts |

Table 8: Overview of the Qiuniu dataset. The upper section shows the distribution of performances based on technical proficiency (dimensions 1–4) and emotional expressiveness (dimension 5), while the lower section summarizes the dataset scale and annotation setup.

- 2: Mostly complete with several noticeable errors; basic techniques achieved
- 1: Frequent interruptions, missing sections, or incorrect techniques

**(4) Musical Flow & Phrasing**

- 3: Clear musical structure with smooth phrasing and appropriate breathing
- 2: Basic structure present but phrasing unclear or slightly mechanical
- 1: Disjointed phrasing and poor musical continuity

**(5) Musical Expression**

- 3: Strong expressive dynamics and emotional engagement
- 2: Basic dynamic variation but limited expressiveness
- 1: Flat dynamics and lack of emotional expression

### B.2 Score-Level Intent Annotation Protocol

All performance videos are annotated by **two expert instructors** with formal training in Chinese traditional music education. Both annotators have extensive experience in evaluating student performances under standardized grading criteria. One annotator is a professional pipa performer and lecturer at a conservatory-level institution, with substantial experience in performance, teaching, and formal music assessment. The other annotator has comparable training and experience in Chinese traditional music education. Each video is evaluated along the five dimensions defined in the TCMS rubric: (1) Pitch & Rhythm, (2) Tone Quality, (3) Completeness, Accuracy & Technique, (4) Musical Flow & Phrasing, and (5) Musical Expression.

Annotations are conducted **independently**, with each annotator assigning scores without access to the other's judgments. A 3-point ordinal scale is used for each dimension, following the rubric described in Appendix B.1. After independent annotation, disagreements are resolved through a **discussion and adjudication process**. Annotators review inconsistent cases jointly and reach a consensus score. This procedure ensures both consistency and adherence to expert-defined evaluation standards. A detailed annotation guideline, based on official grading criteria used in Chinese music education, is provided to annotators prior to the annotation process.

### B.3 Annotation Process and Interface

The annotation process follows a structured evaluation workflow. Annotators first watch each performance video in full to form an overall impression. If necessary, they may re-watch specific segments before assigning scores. Each video is evaluated sequentially across the five TCMS dimensions. Annotators assign scores independently using an online spreadsheet-based interface (Tencent Docs), which allows structured input and efficient data collection. The annotation interface presents each video alongside scoring fields for the five dimensions. A screenshot of the annotation interface is shown in Figure 3.

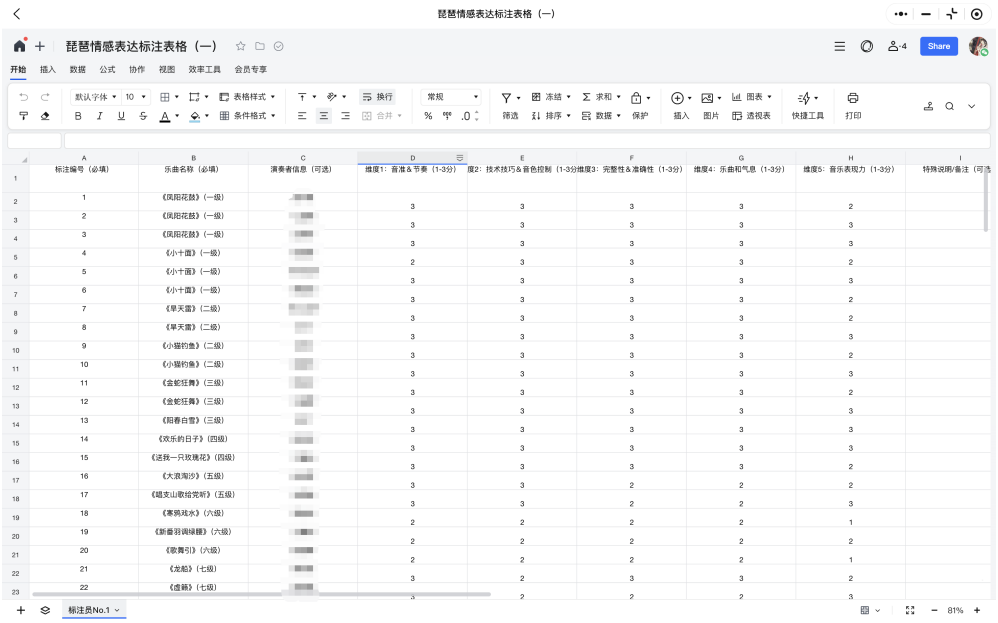


Figure 3: Annotation interface used for scoring performance videos.

# C Evaluation Setup

## C.1 Prompt

To ensure a consistent evaluation protocol across all models, we use a standardized prompt template that instructs the model to evaluate each performance according to the TCMS evaluation framework used by expert annotators. The prompt specifies the evaluation task, the five scoring dimensions, and a 3-point ordinal scale with detailed scoring criteria. Models are required to return structured outputs in JSON format to facilitate automatic parsing and evaluation. The full prompt template used in our experiments is shown in Figure 4.

# D Full Experimental Results

## D.1 Per-Dimension Results

Table 4 reports per-dimension exact accuracy, within-one agreement, $\kappa$, and quadratic-weighted $\kappa$ for all six models across all five evaluation dimensions. Macro-average results are provided in the main paper (Table 2).

# E Repeatability and Case Studies

## E.1 Failure Patterns

Table 9 summarizes the dominant failure pattern and expert verification status observed for each model across three repeated runs.

Table 9: Observed failure patterns and expert verification status.

| Model | Main Pattern | Check |
|---|---|---|
| SALMONN-7B | Score collapse; fabricated evidence | Confirmed |
| Qwen-Omni | High-score collapse; timestamp drift | Confirmed |
| MiMo-V2-Omni | Run-to-run instability | Case-level |
| Gemini-2.5-Pro | Instability across all metrics | Inconclusive |
| Phi-4 | Placeholder timestamps | Pattern only |
| MiniCPM-o-4.5 | Stable high-score collapse | Pattern only |

## E.2 Case Studies

Table 10 provides an overview of the three representative cases with expert verification verdicts. Detailed analyses follow below.

**Case 1: Evidence hallucination (SALMONN, *Chen Xingyuan Hefan*, ID 99).** Both expert annotators assign perfect scores (3/3/3/3/3) to this Level-8 performance. SALMONN stably produces 2/2/2/2/2 across all three runs, citing as evidence for its *Completeness / Accuracy / Technique* score that "the performer makes adjustments to the instrument" at 00:10–00:15. The music expert's verdict is unambiguous: "No tuning occurred." The performer did not adjust the instrument at the cited timestamp; the model fabricated a non-musical event and treated it as grounds for a reduced technical score. This case directly confirms that high score repeatability (100% in Table 7) does not imply reliable grounding: the model consistently cites the same hallucinated evidence across runs.

Table 10: Representative case studies with expert verification. Scores are five-dimensional vectors (P&R / TC / CAT / MF / Expr). Expert Verdict reports the music expert's assessment of the model's claim validity.

| Case | Piece | Model | Expert | Model | Model Claim | Expert Verdict |
|---|---|---|---|---|---|---|
| Hallucination | *Chen Xingyuan* (Lv.8) | SALMONN | 3/3/3/3/3 | 2/2/2/2/2 ×3 | "Performer makes adjustments to the instrument" [00:10–00:15] cited as evidence for reduced technical score | **Incorrect.** "No tuning occurred." |
| Systematic over-estimation | *Han Tian Lei* (Lv.2) | Qwen-Omni | 1/1/1/1/1 | 3/3/3/3/3 ×3 | "Accurate pitch, steady rhythm, fluent technique, near-perfect performance" [00:10–01:30] | **Incorrect.** "Performance interrupted midway; completeness insufficient." |
| Fabricated expressivity | *Jin She Kuang Wu* (Lv.3) | Qwen-Omni | 1/1.5/2/1/1 | 3/3/3/3/3 ×3 | "Energy and joy through energetic articulation; strong emotional engagement" [00:04–01:50] | **Incorrect.** "Errors present; tempo completely different from requirements; no expressivity or emotional engagement." |

**Case 2: Systematic overestimation (Qwen-Omni, *Han Tian Lei*, ID 52).** Expert annotators rate this Level-2 performance at 1/1/1/1/1 across all dimensions. Qwen-Omni produces 3/3/3/3/3 in all three runs, describing pitch as "accurate," rhythm as "steady and consistent," and the overall performance as near-perfect. The expert verdict: "The model's judgment is completely inconsistent with the actual performance; the performance is interrupted midway and completeness is insufficient." This case illustrates that Qwen-Omni's stable high-score collapse produces confident, fluent descriptions that are directly contradicted by the actual recording. Crucially, the overestimation is not confined to *Musical Expressivity*—the model fails equally on objective technical dimensions including completeness and pitch accuracy.

**Case 3: Fabricated expressivity claim (Qwen-Omni, *Jin She Kuang Wu*, ID 49).** Expert scores are 1/1.5/2/1/1, with particularly low marks on *Musical Expressivity* (1). Qwen-Omni assigns 3/3/3/3/3, claiming the performance conveys "energy and joy through energetic articulation" with "strong emotional engagement." The expert's verdict: "Errors present; tempo completely different from requirements; no expressivity or emotional engagement whatsoever." The specific expressivity failure—tempo so wrong that the intended festive character is entirely destroyed—goes entirely undetected by the model, which instead generates a generic positive description of expressivity.

Taken together, all six expert-verified cases received an "Incorrect" verdict, confirming that model-generated evidence does not reliably correspond to actual musical content. Model outputs on Qiuniu are not grounded in musical content regardless of whether the dimension is technical or expressive, and high-confidence outputs can be systematically wrong even on the most objectively assessable dimensions.

## TCMS Evaluation Prompt Template for Pipa Performance

**You are a professional examiner of Chinese traditional music performance.**

You will evaluate a pipa performance video according to the TCMS evaluation framework used by expert annotators.

---

**Evaluation Framework (TCMS)**

Evaluate the performance on 5 dimensions, each scored 1–3:
3 = Excellent
2 = Acceptable
1 = Poor

---

**Scoring Criteria:**

**1. Pitch and Rhythm Accuracy**
3: Accurate pitch, stable rhythm, correct complex patterns.
2: Generally correct, minor pitch deviations, rhythm mostly correct.
1: Frequent pitch errors, unstable rhythm.

**2. Tone Control**
3: Pure, traditional, characteristic tone.
2: Acceptable but noisy or dry.
1: Obvious noise or thin, poor tone.

**3. Completeness, Accuracy and Technical Skill**
3: Smooth, no mistakes, complete, accurate techniques.
2: Several noticeable mistakes, basic techniques achieved.
1: Frequent interruptions or major errors.

**4. Musical Flow and Breathing**
3: Clear phrasing, smooth transitions.
2: Structure complete but unclear phrasing.
1: Disorganized phrasing, broken flow.

**5. Musical Expressiveness**
3: Strong dynamics, clear emotional expression.
2: Basic dynamics, limited emotional depth.
1: Mechanical, little expression.

---

**Instructions:**
- Evaluate how well the performer conveys the INTENDED EMOTION.
- Strictly follow the rubric.
- Do not add extra explanations.
- Return ONLY valid JSON.

---

**Output JSON format:**

```
{
  "pitch_rhythm": 1-3,
  "tone_control": 1-3,
  "completeness_accuracy_technique": 1-3,
  "musical_flow_breathing": 1-3,
  "musical_expressiveness": 1-3,
  "remarks": "short comment if needed"
}
```

Figure 4: Prompt template used to evaluate pipa performances in Qiuniu.