Can Multimodal LLMs Evaluate Music Performance?

Anonymous ACL submission

Abstract

Understanding emotion in music provides a distinctive testbed for probing whether multimodal large language models (MLLMs) can infer intended meaning from behavior rather than rely on explicit verbal descriptions. Unlike text-based Theory of Mind (ToM) evaluations, vocal- and lyric-free music removes the linguistic channel and requires models to judge expressive intent from implicit audio-visual cues. Yet existing benchmarks in ToM, music understanding, and multimodal reasoning largely emphasize explicitly stated mental states, directly observable affective signals, or surface-level musical correlates, leaving intended-emotion inference insufficiently evaluated. We introduce **Qiuniu**, the first benchmark grounded in Traditional Chinese Music Scoring (TCMS) for evaluating intended-emotion inference in vocal- and lyric-free music performance. Qiuniu contains lyric-free *pipa* performance videos paired with score-level annotations of intended emotion and imagery, together with a five-dimension expert rubric that separates technical competence from expressive communication. This design enables evaluation of whether a model can judge not only how well a performance is executed, but also whether it conveys the intended expressive meaning of the piece. Experiments show that current MLLMs achieve relatively strong agreement with experts on technical dimensions, but fall substantially behind on *Musical Expressivity*, the dimension most directly tied to intended meaning. This structured gap suggests that current models can extract surface-level performance cues, yet remain limited in mapping multimodal behavior to expressive intent. Our results position Qiuniu as a benchmark for intent-level multimodal understanding and a culturally grounded test of ToM-like inference in music.

1 Introduction

Can multimodal large language models (MLLMs) infer the emotion a performer intends to convey in vocal- and lyric-free music performance, as humans often can? The target here is not an acoustic descriptor but intended emotion, a latent mental state, so solving it requires a minimal Theory of Mind (ToM) step, reading intention from multimodal behavior rather than labeling sound and visual patterns. This is a necessary bar for intent-level multimodal understanding, because vocal- and lyric-free music removes the verbal channel entirely, leaving no words to lean on, while still offering strong shortcuts in both sound and sight, such as tempo, loudness, articulation, and even exaggerated gestures that can look right for the wrong reasons. With controlled cues or cue conflicts across audio and video, the task cleanly exposes whether an MLLM treats a performance as meaningful communicative behavior or as a bundle of correlates. Traditional Chinese Music Scoring (TCMS) operationalizes this by construction. Each canonical piece is taught with expert-defined emotion and imagery, and grading asks whether the performance conveys that intended meaning under restrained expression. The exemplar High Mountains and Flowing Water (at least 400 BC) states the goal plainly, Boya plays the qin without words and Ziqi hears high mountains then flowing water, while the counter idiom dui niu tan qin, literally playing the lute to a cow, names the opposite, hearing and seeing yet missing the meaning, like models that follow surface correlates. Beyond benchmarking, China's grading exams already handle roughly 1.57 million registrations, so even modest, reliable triage could translate into substantial expert-time savings.

Most MLLM benchmarks do not isolate ToM-style intent inference from multimodal behavior performance, because the target is typically available from explicit language or other directly readable channels, and the evaluation rarely forces models to infer what an agent meant from how they performed. Multimodal emotion and empa-

thy benchmarks largely emphasize directly observable signals such as text, facial expressions, or speech (Lian et al., 2023; Zadeh et al., 2018). ToM evaluations are predominantly text-based, from false-belief tasks (Baron-Cohen et al., 1985; Strachan et al., 2024) to advanced belief-reasoning paradigms (Happé, 1994; van Duijn et al., 2023), where mental states are stated in language rather than inferred from behavior. Music emotion benchmarks (e.g., MusCaps; MuChoMusic) likewise focus on recognizing perceived affect from salient features, often including lyrics or relying on coarse correlates such as tempo (Manco et al., 2021; Weck et al., 2024).

To address it, we propose Qiuniu, the first benchmark grounded in Traditional Chinese Music Scoring (TCMS) that pairs vocal- and lyric-free performance videos with score-level intent annotations and expert rubrics, enabling an evaluation of MLLMs' ability to infer a performer's intended emotion. Qiuniu contains N pipa performance videos spanning M canonical pieces, where each video is aligned with annotated sheet music that specifies the intended emotion and imagery, and the collection includes both emotionally expressive and deliberately restrained renditions to reduce reliance on obvious surface cues. To obtain reliable supervision, we co-design a K-dimension TCMS scoring rubric with TCM educators and computer scientists, with emotional expressivity as the primary dimension, and collect multi-expert ordinal ratings on a 1–3 scale. We report inter-expert agreement to establish the reliability of human judgments. Based on this setup, we define the task as intent-grounded performance grading that, given a performance video and its score-level intent annotation, a model assigns the same K-dimension scores by judging whether the performance conveys the intended meaning, rather than exploiting audio–visual shortcuts such as tempo, loudness, articulation, or exaggerated gestures. We evaluate model–expert alignment using exact accuracy on the ordinal 1–3 ratings, and additionally report within-one-level agreement (||1) to reflect near-miss consistency, both are reported per dimension and macro-averaged across dimensions.

In our evaluation of M representative MLLMs, spanning closed and open models, general-purpose and multimodal-specialized systems, and a wide range of sizes—we reveal a fundamental limitation in intent-level multimodal understanding on Qiuniu. Despite strong performance on standard multimodal benchmarks, even top-tier models struggle to grade emotional expressivity in vocal- and lyric-free Traditional Chinese music performance when the target is the score-specified intended emotion and imagery, and model–expert agreement remains well below the expert–expert reliability ceiling. The gap is not a generic music understanding issue. When we hold the input and rubric fixed but switch the target, models score technical execution far more consistently than expressive intent, showing that the dominant failure lies in the ToM-like step of inferring what the performer meant to convey, not in basic perception of audio–visual cues. This diagnosis is reinforced by controlled evaluations. Improvements from audio-only or video-only to audio and video are limited, and under cue-controlled or cue-conflict conditions models systematically drift toward salient heuristics rather than staying anchored to intended meaning. Scaling model size offers only partial relief and quickly exhibits diminishing returns, indicating that more parameters alone do not resolve intent inference. Overall, Qiuniu exposes a bottleneck that persists even for strong MLLMs. Translating multimodal behavior into intended emotion remains fragile precisely when surface cues are predictive but not decisive.

Our main contributions are:

(1) We introduce Qiuniu, the first TCMS-grounded benchmark for evaluating MLLMs on intended-emotion inference from vocal- and lyric-free Traditional Chinese music performances, pairing audio–video performances with score-level intent so that success cannot be driven by verbal cues.

(2) We release the Qiuniu dataset together with its TCMS grading rubric: N pipa performance videos aligned with culturally annotated sheet music, and a five-dimension scoring framework co-designed with TCM experts. Using two expert raters, we obtain reliable ordinal (1–3) annotations and report inter-rater consistency as the benchmark's reference standard.

(3) We systematically evaluate multimodal, acoustic-only, and symbolic-only models across scales and training paradigms, and provide diagnostic evidence that current models lag behind experts on emotional expressivity even when they score technical aspects more reliably, highlighting how modality access, scale, and data diversity shape alignment with human judgments.

2 Related Work

2.1 Theory of Mind Evaluation in Language Models

Theory of Mind (ToM) originates from psychology and concerns the ability to attribute beliefs, intentions, and other mental states to others (Baron-Cohen et al., 1985; Happé, 1994). Recent evaluations of ToM in large language models have largely relied on text-based benchmarks, including false-belief tasks (Strachan et al., 2024) and more advanced belief-reasoning settings (van Duijn et al., 2023; Chen et al., 2024; Kosinski, 2024). In these benchmarks, the relevant mental states are typically specified explicitly in language, for example through nested statements such as one person believing that another person thinks someone is happy. As a result, success primarily reflects reasoning over stated propositions rather than inferring an agent's intended meaning from behavior. Our setting differs in a crucial way. The target is not an explicitly described belief state, but the intended expressive meaning conveyed through performance.

2.2 Music Emotion and Multimodal Emotion Benchmarks

A related line of work studies emotional understanding in music and multimodal data, but most existing benchmarks emphasize recognition rather than intent-grounded inference. In music benchmarks such as MusCaps and MuChoMusic, models are asked to predict mood, expression, or related affective properties from musical content (Manco et al., 2021; Weck et al., 2024). These tasks often center on perceived affect or salient musical correlates, such as tempo, harmony, instrumentation, or timbre. Prior analyses further suggest that some systems can exploit linguistic or other superficial cues without relying deeply on the musical signal itself (Weck et al., 2024). Such benchmarks therefore do not cleanly isolate whether a model can recover the intended meaning behind a performance. Similarly, multimodal emotion and empathy benchmarks, such as CMU-MOSEI and MER 2023, mainly evaluate models on directly observable affective signals from text, facial expressions, and speech (Zadeh et al., 2018; Lian et al., 2023). These settings are important for affect recognition, but they do not require models to infer what an agent means to communicate when the target is not directly stated or visibly manifest. A structured comparison between our benchmark and existing ToM and emotion understanding benchmarks is summarized in Table 1.

3 Qiuniu

3.1 Task Definition

Qiuniu is grounded in *Traditional Chinese Music Scoring* (TCMS), a performance assessment practice widely used in Chinese music pedagogy and graded examinations. Crucially, TCMS does not evaluate performance solely in terms of technical correctness. Instead, it combines two kinds of judgment: whether the performer executes the piece accurately and coherently, and whether the performance conveys the emotion, imagery, and artistic conception associated with the piece. This makes TCMS a natural setting for studying *intent-grounded performance evaluation*, where the goal is not only to determine whether a performance is correct, but also whether it communicates the intended expressive meaning of the score.

To convert this pedagogical practice into a benchmark task, we operationalize TCMS with a five-dimensional rubric that separates technical execution from expressive communication while preserving their interaction in real evaluation. The first four dimensions capture whether the performance is musically well-formed. *Pitch & Rhythm* measures basic correctness and stability in intonation and timing, since major deviations at this level directly undermine the identity of the piece. *Tone Control* evaluates the quality and consistency of sound production, including clarity, steadiness, and control, which are essential because the same notes can be rendered with markedly different sonic quality. *Completeness, Accuracy & Technical Proficiency* assesses whether the required musical material is fully performed, whether the realization matches the score, and whether the performer demonstrates the instrumental techniques expected by the piece. *Musical Flow & Breath Control* captures structural coherence: whether phrases connect naturally, whether the musical line unfolds continuously, and whether the performance maintains an appropriate sense of pacing and breathing. Together, these four dimensions describe whether a performance is technically and musically competent.

The fifth dimension, *Musical Expressivity*, serves a different role. Rather than asking whether the performance is merely correct, it asks whether the performer successfully conveys the intended

Benchmark	Domain	Modality	Task	TI	IS	AS	Target
False-belief Tasks (Strachan et al., 2024)	Social reasoning	Text	QA	✓	✗	✗	Belief reasoning
ToMBench (Chen et al., 2024)	Social reasoning	Text	MCQ	✓	✗	✗	Nested belief
MusCaps (Manco et al., 2021)	Music	Audio	Captioning	✗	✗	✗	Music caption
MuChoMusic (Weck et al., 2024)	Music	Audio+Text	QA / Cls	✗	✗	✗	Emotion / attribute
CMU-MOSEI (Zadeh et al., 2018)	Multimodal affect	V+A+T	Sentiment	✗	✗	✗	Sentiment prediction
MER2023 (Lian et al., 2023)	Multimodal affect	V+A+T	Emotion	✗	✗	✗	Emotion recognition
Qiuniu-Bench (Ours)	Music performance	Video+Audio	Rubric eval	✓	✓	✓	**Expressive intent**

Table 1: Comparison between Qiuniu-Bench and existing benchmarks. **TI** denotes **ToM Inference**, indicating whether the task evaluates Theory of Mind reasoning rather than simple affect recognition. **IS** denotes **Implicit Scenario**, where the mental state is not explicitly stated and must be inferred from context. **AS** denotes **Action-based Signal**, indicating that inference relies on observed behavior or interactions rather than explicit linguistic description.

emotional meaning and artistic imagery of the piece. This dimension is therefore the primary target of interest in Qiuniu. Importantly, it cannot be reduced to any single observable correlate such as tempo, loudness, or visible gesture. A performance may be technically mature yet emotionally restrained, or it may display salient expressive cues without faithfully communicating the intended meaning of the score. For this reason, *Musical Expressivity* is treated as an intent-grounded evaluative dimension rather than a direct emotion label.

Formally, each sample in Qiuniu consists of a vocal- and lyric-free performance video X_i, a score-level intent annotation S_i, and the shared TCMS rubric R. The intent annotation specifies the intended emotion and imagery associated with the performed piece, as defined by the teaching and scoring tradition. Given (X_i, S_i, R), the model predicts a five-dimensional ordinal rating vector

$$Y_i = \left(y_i^{(1)}, y_i^{(2)}, y_i^{(3)}, y_i^{(4)}, y_i^{(5)}\right), \qquad (1)$$

where each $y_i^{(d)} \in \{1, 2, 3\}$ denotes the score for one rubric dimension. We denote the dimensions in order as *Pitch & Rhythm*, *Tone Control*, *Completeness, Accuracy & Technical Proficiency*, *Musical Flow & Breath Control*, and *Musical Expressivity*.

Under this formulation, Qiuniu is not a standard music emotion recognition task. The target is not perceived affect, nor is it a surface acoustic descriptor. Instead, the model must judge how well a performance realizes the intended expressive meaning of the piece within a real grading framework. That is, the model must map from multimodal performance behavior to expert-style ordinal judgments, especially on the *Musical Expressivity* dimension, where success depends on evaluating alignment between the performed realization and the score-conditioned intent.

Each sample is independently rated by two expert annotators under the same rubric, yielding reference labels

$$Y_i^{(A)} \quad \text{and} \quad Y_i^{(B)}. \qquad (2)$$

Model predictions are compared against these expert ratings to measure alignment with human judgment. Because all labels are ordinal, Qiuniu is naturally formulated as a multi-dimensional ordinal prediction task.

3.2 Dataset Curation and Annotation

Data Source and Collection. We construct Qiuniu in partnership with musical schools in China. The dataset consists of vocal- and lyric-free *pipa* performance videos drawn from graded examination repertoire spanning from Levels 1 to 11. In total, Qiuniu contains 49 performance videos covering 24 canonical pieces. Each video is paired with the corresponding sheet music and its score-level annotation of intended emotion and imagery. The dataset covers students with varying proficiency levels and includes both emotionally expressive renditions and deliberately restrained performances. This design reduces the possibility of solving the task through trivial shortcuts, such as uniformly assigning low scores to weak performances or relying only on salient cues such as loudness, tempo, or exaggerated body movement. Importantly, the dataset also includes cases in which technical execution is relatively mature while emotional communication remains limited. An overview of the dataset statistics and performance distribution is shown in Table 2.

Score-Level Intent Annotation. A central feature of Qiuniu is that each performance is anchored

Performance Distribution	Definition	# Videos
Technical strong, Emotion weak	Tech $\geq$2.5 & Emotion $\leq$2	14
Emotion strong, Technical weak	Emotion $\geq$2.5 & Tech $\leq$2	4
Both strong	Tech $\geq$2.5 & Emotion $\geq$2.5	18
Both weak	Tech $\leq$2 & Emotion $\leq$2	13
Total videos		49
Unique musical pieces		24
Total duration		198.0 min ($\approx$3.3 hours)
Average duration per sample		242.5 s ($\approx$4.0 min)
Difficulty levels		1–10
Annotation dimensions		5
Score scale		1–3
Annotators		2 experts

Table 2: 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.

to *score-level intent* rather than to a free-form or post hoc emotion label. To convert this evaluative practice into benchmark supervision, we develop a standardized five-dimension rubric grounded in TCMS and co-designed by computer science researchers and experienced Traditional Chinese Music educators. The rubric covers both technical and expressive aspects of performance, following our task design, including *Pitch & Rhythm*, *Tone Control*, *Completeness, Accuracy & Technical Proficiency*, *Musical Flow & Breath Control*, and *Musical Expressivity*. A detailed annotation guideline (Appendix I) was developed iteratively through multiple rounds of pilot annotation and discussion, with concise scoring criteria provided for each dimension to ensure consistent interpretation. All samples were independently rated by two domain experts, both pipa instructors with more than ten years of teaching experience and experience serving as certified music examination judges. For each dimension, experts assign an ordinal score from 1 to 3, where higher scores indicate stronger performance under the corresponding criterion. These expert ratings serve as the benchmark reference labels for model evaluation.

Annotation Reliability. We assess inter-rater reliability using three complementary metrics, including exact-match Accuracy, quadratic-weighted Cohen's κ, and Gwet's AC2. Accuracy measures direct agreement, quadratic-weighted κ accounts for the ordinal nature of the scale while penalizing larger disagreements more heavily, and Gwet's AC2 provides a more stable estimate under potential category imbalance. Across the five dimensions, the two experts exhibit strong overall agreement, with 0.98 Accuracy, 0.98 quadratic-weighted κ, and 0.97 Gwet's AC2 (Table 3). Most dimensions achieve perfect or near-perfect consistency, while the few disagreements occur primarily in technically adjacent cases and typically remain within one rating level. These results support the reliability of the expert annotations and justify their use as reference judgments in Qiuniu.

Table 3: Inter-annotator Agreement Metrics Across Five Dimensions

Dimension	Accuracy	κ	AC2
Pitch & Rhythm	1.000	1.000	1.000
Technical Skill	0.959	0.947	0.928
Completeness	0.959	0.948	0.927
Musical Flow	1.000	1.000	1.000
Expressivity	1.000	1.000	1.000
Macro Average	**0.984**	**0.979**	**0.971**

3.3 Evaluation Protocol

For each evaluation instance, a model is given three inputs: (1) a vocal- and lyric-free performance video, (2) the score-level intent annotation specifying the intended emotion and imagery of the piece, and (3) the five-dimension TCMS rubric. To ensure comparability across systems, we use a standardized prompt template for all models. Rather than asking models to freely describe the performance or classify its perceived emotion, the prompt asks them to evaluate the performance under the same rubric used by expert annotators and to assign one ordinal score in $\{1, 2, 3\}$ for each dimension. The prompt further instructs models to follow the rubric strictly, avoid unsupported speculation, and keep scoring standards consistent across samples.

A brief *Remarks* field is allowed for unusual interpretive or technical cases, but every model must output a concrete score for all five dimensions. The full prompt template is provided in Appendix C.

> You are given a performance video, the intended emotion and imagery of the piece, and the TCMS scoring rubric. Evaluate the performance under this rubric and assign one score from 1–3 for each of the following dimensions: (1) Pitch & Rhythm, (2) Tone Control, (3) Completeness, Accuracy & Technical Proficiency, (4) Musical Flow & Breath Control, and (5) Musical Expressivity. Base your judgment on whether the performance realizes the intended meaning of the piece, keep the scoring standard consistent, and optionally provide brief remarks for special cases.

Each sample in Qiuniu is independently rated by two expert annotators under the same rubric. We evaluate model-expert alignment by comparing model predictions against the expert reference ratings and averaging results across the two annotators unless otherwise stated. Because Qiuniu is a multi-dimensional ordinal rating task, we use quadratic-weighted Cohen's κ as the primary metric, since it accounts for chance agreement and penalizes larger rating discrepancies more heavily. We additionally report exact-match Accuracy as a complementary measure of strict score agreement, and within-one-level agreement to capture near-miss consistency on the 1 to 3 scale. All metrics are reported for each individual dimension and as macro-averages across the five dimensions. Expert-expert agreement is reported separately as a human reliability reference for interpreting model performance.

4 Experiments

4.1 Experimental Settings

We evaluate a diverse set of representative MLLMs, including both closed-source API systems and open-source models. Our selection spans general-purpose multimodal models developed for broad visual–language reasoning, as well as open-source vision-language models designed for scalable multimodal understanding (OpenAI et al., 2024; Team et al., 2025; Tang et al., 2024). These systems vary substantially in architecture, training data, alignment strategy, and parameter scale, allowing us to examine whether performance on Qiuniu depends primarily on model family, modality design, or scale. A full list of evaluated models is provided in Table 4. For open-source models, we use their official inference frameworks; for closed-source systems, we access them through their corresponding APIs. Unless otherwise stated, all models are evaluated under the same prompt template, scoring instructions, and output format described in Appendix C.

Model	Family	Params	Modality	Access
GPT-4o	OpenAI	–	Audio + Video	API
Gemini 1.5 Pro	Google	–	Audio + Video	API
Salmonn-7B	SALMONN	7B	Audio + Video	Local

Table 4: Audio-visual multimodal models evaluated in our experiments.

4.2 Main Results

Table 5 summarizes model performance on Qiuniu. Across all evaluated systems, even the strongest MLLMs remain substantially below expert-expert agreement. The best-performing model, Salmonn-7B, achieves a macro-average quadratic-weighted Cohen's κ of -0.024 and an exact-match Accuracy of 0.075, whereas expert-expert agreement reaches substantially higher levels. Because human agreement is high, this gap is not best understood as annotation noise or task instability; rather, it indicates that Qiuniu captures a capability that current MLLMs still do not reliably master.

More importantly, this gap is highly structured rather than uniform across dimensions. Models perform relatively better on the four primarily technical dimensions, including *Pitch & Rhythm*, *Tone Control*, *Completeness, Accuracy & Technical Proficiency*, and *Musical Flow & Breath Control*, than on *Musical Expressivity*. For Salmonn-7B, the average κ across the four technical dimensions is -0.01, compared with only -0.081 on *Musical Expressivity*, yielding a gap of 0.071. If Qiuniu were difficult mainly because musical performance is challenging to perceive audiovisually, model performance would be expected to degrade more evenly across dimensions. Instead, the selective weakness on *Musical Expressivity* localizes the main bottleneck to judging whether the performed behavior successfully conveys the score-specified intended emotion and imagery. The fact that current MLLMs remain comparatively stronger on technical competence yet substantially weaker on *Musical Expressivity* suggests that they can extract many low-level and mid-level performance cues, but still struggle at the final interpretive step of mapping those cues to intended meaning.

The ranking across model families leads to the

same conclusion. Larger and more capable general-purpose MLLMs generally achieve stronger overall performance, indicating that scale and stronger multimodal pretraining are helpful. However, these gains are only partial, improvements are more consistent on technical dimensions than on *Musical Expressivity*, and even the strongest systems do not approach expert reliability on the latter. Taken together, the results suggest that Qiuniu exposes a specific weakness in current MLLMs. They are considerably better at judging whether a performance is technically well formed than at determining whether it successfully communicates the intended affective and artistic meaning of the piece.

4.3 Technical Competence vs. Expressive Intent

The main results suggest that model errors on Qiuniu are not uniformly distributed across evaluation dimensions. To better localize the source of failure, we compare model performance on the four primarily technical dimensions—*Pitch & Rhythm*, *Tone Control*, *Completeness, Accuracy & Technical Proficiency*, and *Musical Flow & Breath Control*—with performance on *Musical Expressivity*. Across model families, agreement with expert ratings is consistently higher on the technical dimensions than on *Musical Expressivity*. This pattern indicates that current MLLMs can capture a substantial amount of information about low-level and mid-level performance quality, including correctness, stability, fluency, and control, yet remain less reliable when asked to judge whether a performance conveys the intended emotion and imagery of the piece.

This distinction is central to the interpretation of Qiuniu. If the benchmark were difficult mainly because musical performance is challenging to perceive from audio and video, one would expect broadly similar degradation across all five dimensions. Instead, the observed technical–expressive asymmetry points to a more specific bottleneck. Current models appear more capable of evaluating whether a performance is musically well formed than of determining whether it realizes the score-conditioned intended meaning required by the TCMS framework.

Importantly, *Musical Expressivity* is not simply a noisier version of technical scoring. Under TCMS, it requires the evaluator to assess whether expressive choices are appropriate for the intended affective and artistic character of the piece, rather than merely whether salient cues such as louder dynamics, faster tempo, or more visible gesture are present. The fact that model performance drops most sharply on this dimension therefore suggests a limitation in intent-grounded evaluation rather than a general failure of multimodal perception.

Taken together, these results support the central claim of the paper: current MLLMs are substantially stronger at assessing technical competence than at inferring expressive intent from multimodal musical behavior. In other words, the main challenge exposed by Qiuniu lies not in recognizing what happens in a performance, but in judging what the performance is meant to convey.

4.4 Case Studies

To better understand where current MLLMs fail on Qiuniu, we qualitatively examine representative cases of model–expert disagreement. Rather than treating these errors as isolated failures, we use them to identify recurring patterns in how models score performances under the TCMS rubric. Across cases, the dominant pattern is not a complete inability to assess performance quality, but a mismatch between surface-level performance cues and intended expressive meaning.

In the first type of case, models assign relatively accurate scores on technical dimensions but overestimate *Musical Expressivity*. These performances are typically well executed in pitch, rhythm, tone, and overall fluency, yet experts judge that they do not fully convey the score-specified emotion and imagery. Such cases suggest that models often conflate technical maturity with expressive success, even though TCMS explicitly distinguishes between the two. This pattern is consistent with our quantitative finding that technical alignment is substantially stronger than expressive alignment.

In the second type of case, models appear to rely excessively on salient surface cues such as stronger dynamics, faster tempo, or visually expressive gestures. When these cues are present, models tend to assign higher expressive scores, even when expert raters judge that the performance does not match the intended affective character of the piece. Qualitatively, this suggests that current MLLMs are sensitive to observable correlates of expressiveness, but do not always evaluate whether those correlates are appropriate for the score-conditioned intended meaning.

In the third type of case, the model's explana-

Model	Macro κ	Tech. Avg.	Express.	Pitch & Rhythm	Tone	Technique	Musical Flow	Gap
Salmonn-7B	-0.024	-0.01	-0.081	-0.084	0.000	0.000	0.044	0.071
Expert-Expert	0.979	0.974	1.000	1.000	0.947	0.948	1.000	-0.026

Table 5: Quadratic-weighted Cohen's κ on Qiuniu.

tion correctly identifies local performance properties but still leads to an expert-misaligned final judgment. For example, a model may note stable rhythm, clear tone, or restrained delivery, yet fail to translate those observations into an appropriate score on *Musical Expressivity*. These cases indicate that the main difficulty is not only perceptual, but interpretive: the model can often describe what happens in the performance, but struggles to determine what the performance is meant to convey under the TCMS framework.

Taken together, these case studies reinforce the main conclusion of the paper. The central weakness of current MLLMs on Qiuniu is not simply poor multimodal perception, but an unstable mapping from multimodal performance behavior to intended expressive meaning.

5 Conclusion

In this paper, we introduce Qiuniu, the first benchmark grounded in Traditional Chinese Music Scoring (TCMS) for evaluating multimodal large language models' ability to infer intended emotions from implicit musical expression. This gap suggests that current models rely primarily on surface-level perceptual signals and remain limited in Theory-of-Mind-like reasoning when inferring human emotional intent. Qiuniu provides a diagnostic benchmark for studying implicit emotional inference in music performance and offers a foundation for future research on improving multimodal models' alignment with human expressive judgments.

6 Limitations

Despite these contributions, our work has several limitations: **(1) Instrument Scope.** The Qiuniu dataset focuses exclusively on pipa performances, as the benchmark is designed around a specific instrumental tradition. This focus may limit the generalizability of our findings to other musical instruments or performance contexts. As a result, the evaluation of multimodal models in this work is also restricted to pipa performance videos. Future work could extend the dataset to include additional instruments and musical traditions to examine whether the observed patterns hold more broadly. **(2) Dataset Size.** Although the dataset is curated and annotated by expert instructors, its overall scale remains relatively limited, consisting of 44 performance videos. Expanding the dataset with a larger number of performances, performers, and repertoire could further improve the robustness and statistical reliability of future evaluations. **(3) Model Coverage.** Our evaluation currently covers a limited set of multimodal large language models. To balance feasibility and model diversity, we include representative systems from both API-based and open-source families (e.g., GPT-based models, Gemini, and Salmonn). However, the number of evaluated models remains small, and future work could extend the benchmark to a broader range of architectures and training paradigms.

Acknowledgments

References

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gong Yao Jiang, Yaru Cao, Meng Ting Hu, Yung Hwei Lai, Zexuan Xiong, and Min Lie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 15959–15983. Association for Computing Machinery.

Francesca G. E. Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2):129–154.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery.

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

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

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.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. 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.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7–10 on advanced tests. *arXiv preprint*.

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.

Amir Zadeh, Paul Pu Liang, Jonathan Van Briesen, Soujanya Poria, Edmund Tong, 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*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

A 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
- 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 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.

C 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 1.

D 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 A.

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.

E Dataset Details

The dataset consists of **49 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.

F 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.

Prompt Template for TCMS Evaluation

Role

You are a professional examiner of Chinese traditional music performance.

Task

Evaluate a pipa performance according to the TCMS evaluation framework used by expert annotators.

Input

1. A pipa performance video 2. The intended musical emotion of the piece 3. The TCMS evaluation rubric

Evaluation Dimensions

The performance should be evaluated on the following five dimensions:

• Pitch and Rhythm Accuracy • Tone Control • Completeness, Accuracy and Technical Skill • Musical Flow and Breathing • Musical Expressiveness

Score Scale

3 = Excellent 2 = Acceptable 1 = Poor

Instructions

Evaluate how effectively the performer conveys the intended emotion through the performance. Base your judgement only on observable musical cues such as timing, tone quality, phrasing, and dynamics. Strictly follow the scoring rubric and maintain consistent scoring standards across samples.

Output Format

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

Example Output

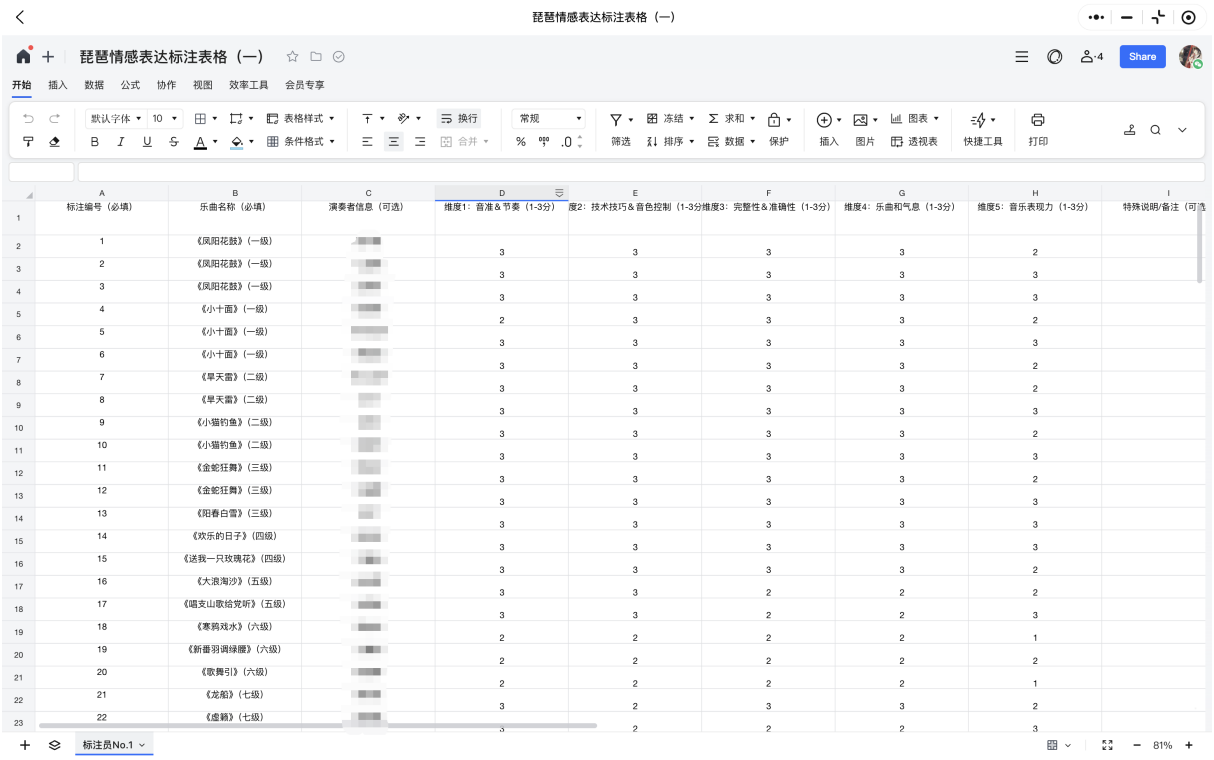
```
{ "pitch_rhythm": 3, "tone_control": 2, "completeness_accuracy_technique": 2, "musical_flow_breathing": 3, "musical_expressiveness": 2, "remarks": "minor rhythm instability" }
```

Figure 1: Prompt template used to evaluate pipa performances in Qiuniu.

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 2.



琵琶情感表达标注表格（一）

标注编号（必填）	乐曲名称（必填）	演奏者信息（可选）	维度1：音准&节奏（1-3分）	度2：技术技巧&音色控制（1-3分	维度3：完整性&准确性（1-3分）	维度4：乐曲和气息（1-3分）	维度5：音乐表现力（1-3分）	特殊说明/备注（可选
1	《凤阳花鼓》（一级）	[redacted]	3	3	3	3	2	
2	《凤阳花鼓》（一级）	[redacted]	3	3	3	3	3	
3	《凤阳花鼓》（一级）	[redacted]	3	3	3	3	3	
4	《小十面》（一级）	[redacted]	2	3	3	3	2	
5	《小十面》（一级）	[redacted]	3	3	3	3	3	
6	《小十面》（一级）	[redacted]	3	3	3	3	2	
7	《旱天雷》（二级）	[redacted]	3	3	3	3	2	
8	《旱天雷》（二级）	[redacted]	3	3	3	3	3	
9	《小猫钓鱼》（二级）	[redacted]	3	3	3	3	2	
10	《小猫钓鱼》（二级）	[redacted]	3	3	3	3	3	
11	《金蛇狂舞》（三级）	[redacted]	3	3	3	3	2	
12	《金蛇狂舞》（三级）	[redacted]	3	3	3	3	3	
13	《阳春白雪》（三级）	[redacted]	3	3	3	3	3	
14	《欢乐的日子》（四级）	[redacted]	3	3	3	3	3	
15	《送我一只玫瑰花》（四级）	[redacted]	3	3	3	3	2	
16	《大浪淘沙》（五级）	[redacted]	3	3	2	2	2	
17	《唱支山歌给党听》（五级）	[redacted]	3	3	2	2	3	
18	《寒鸦戏水》（六级）	[redacted]	2	2	2	2	1	
19	《新垂羽调绿腰》（六级）	[redacted]	2	2	2	2	2	
20	《歌舞引》（六级）	[redacted]	2	2	2	2	1	
21	《龙船》（七级）	[redacted]	3	2	3	3	2	
22	《虞翰》（七级）	[redacted]	3	2	2	2	3	

Figure 2: Annotation interface used for scoring performance videos.