BoYaEval: Evaluating Multimodal Large Language Models on Understanding Ancient Chinese Musical Scores

Anonymous ACL submission

Abstract

Large Multimodal Models (LMMs) excel in general tasks but struggle with specialized, structured cultural symbols. We introduce BoYaEval, the first comprehensive benchmark dedicated to deciphering diverse Ancient Chinese musical notations, including five types of ancient Chinese music notation systems. These systems utilize unique spatial layouts and specialized ideograms to encode pitch and intricate playing techniques. BoYaEval comprises 3,175 high-quality images across these notation styles and establishes a three-tier evaluation: Structural Parsing (symbol recognition), Instructional Translation (technique mapping), and Musical Reasoning (melody derivation). We evaluate 22 leading LMMs. Results indicate that while models perform adequately in basic recognition, they fail in cross-system compositional logic, scoring only around 27% on reasoning tasks. BoYaEval highlights the limitations of current LMMs in processing diverse spatial-symbolic dependencies, bridging the gap between ancient wisdom and modern AI for digitizing intangible cultural heritage.

1 Introduction

The advent of Large Multimodal Models (LMMs), exemplified by GPT-4o (Hurst et al., 2024) and Gemini, has revolutionized the field of document intelligence (Luo et al., 2024; Hu et al., 2024). These models demonstrate remarkable proficiency in parsing standard structured documents, such as mathematical formulas, statistical charts, and modern musical scores (e.g., Western staff notation) (Li et al., 2024a; Ding et al., 2025). However, the ambition of Artificial General Intelligence (AGI) is not merely to process contemporary data but to bridge the temporal gap, decoding the vast repository of human civilization's historical records. Despite their success in general domains, current LMMs face a significant "cultural gap" when confronted with specialized, non-Western, and historical symbolic systems .

We argue that Ancient Chinese Musical Scores represent one of the most challenging frontiers for multimodal understanding. Unlike Western staff notation, which primarily visualizes pitch and duration on a linear timeline, ancient Chinese notations—such as Guqin *Jianzipu* (reduced-character notation), *Gongchepu*, and *Suzipu*—function as complex instructional algorithms. For instance, a single *Jianzipu* character is not a phonetic word but a composite ideogram spatially compressing information about string number, finger position, plucking technique, and tonal modification. As shown in Figure 1, deciphering these scores requires a Large Multimodal Model to perform simultaneous spatial parsing, semantic decompression, and cross-modal reasoning from visual symbols to auditory or gestural intent.

Existing benchmarks for document understanding (e.g., DocVQA (Mathew et al., 2021), MathVista (Lu et al., 2023)) focus heavily on text-heavy or diagrammatic data, neglecting the intricate spatial-symbolic dependencies found in intangible cultural heritage. Consequently, the capability of state-of-the-art models to preserve and digitize these endangered musical traditions remains unknown. To systematically probe the depth of multimodal understanding, we propose a novel three-tier evaluation hierarchy that mirrors the cognitive process of interpreting ancient scores: (1) Structural Parsing challenges models to disentangle the complex spatial layout of composite ideograms, requiring fine-grained visual perception to separate interlocking radicals (e.g., distinguishing the "thumb" technique from the "string seven" position) (2) Instructional Translation assesses cross-modal semantic grounding, where models must map abstract visual symbols to executable physical instructions (e.g., translating a glyph into "Pluck the 7th string with the thumb at the 9th hui") (3)

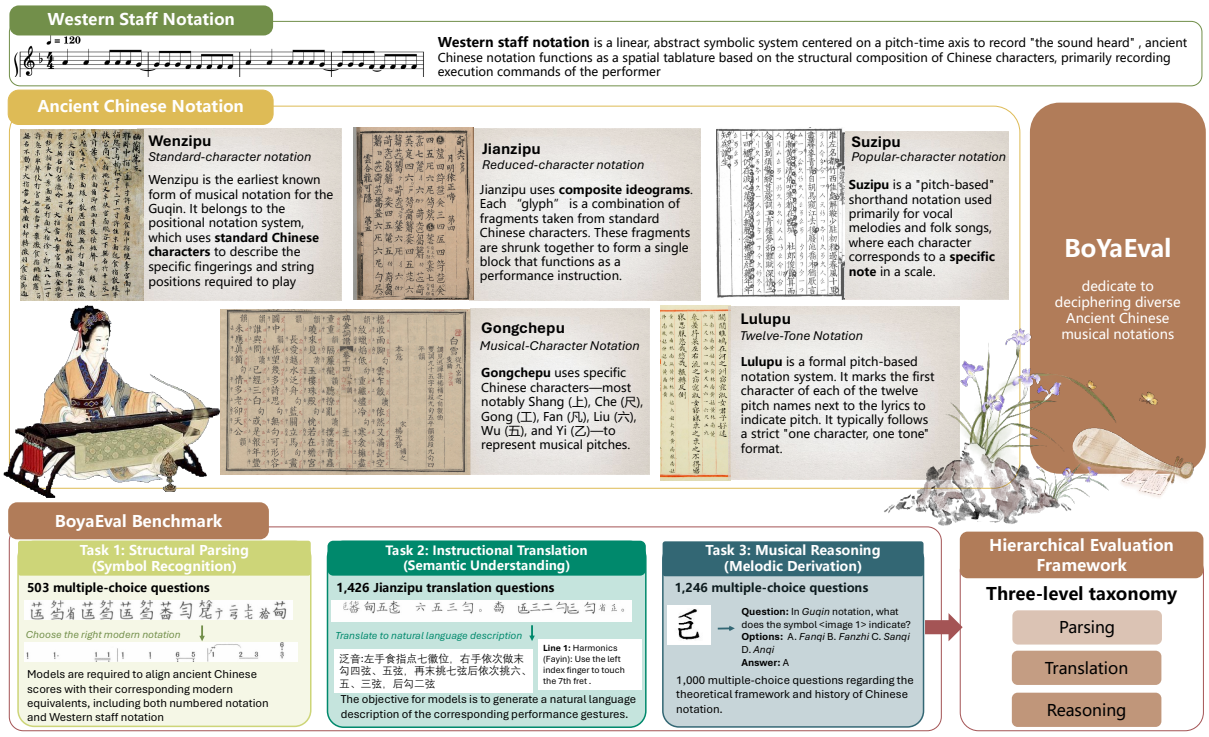

Figure 1: Overview of the **BoyaEval** benchmark. The framework highlights the fundamental differences between Western staff notation and ancient Chinese systems.

Musical Reasoning evaluates high-level logical inference, requiring the synthesis of visual cues with domain-specific musicological knowledge to derive underlying melodies and rhythms.

Experimental results reveal that while current LMMs (e.g., GPT-4o) are competent in surface-level parsing, they exhibit severe deficits in compositional musical reasoning. Addressing these limitations, this work makes the following contributions: (1) Releasing BoYaEval, the first multimodal benchmark for Ancient Chinese Music, containing five types of ancient notations (Jianzipu, Wenzipu, Suzipu, Gongchepu, and Lulupu) (2) Establishing a three-tier evaluation hierarchy to probe symbolic understanding depth (3) Offering a granular diagnosis of spatial reasoning failures in non-linear ideograms (4) Advocating for "AI for Culture" to advance the digitization of intangible heritage

2 BoYaEval: Dataset Construction and Analysis

To rigorously evaluate LMMs on ancient Chinese music understanding, we construct **BoYaEval**, a comprehensive benchmark encompassing diverse notation systems and varying levels of cognitive difficulty. The construction process involves three phases: raw data acquisition, expert annotation, and task formulation.

2.1 Data Collection and Preprocessing

To construct a high-quality corpus, we collaborated with musicologists from the Anonymous Conservatory of Music to curate data from two primary authoritative sources. First, we obtained digitized manuscripts from seminal anthologies, such as the Shen Qi Mi Pu (1425 AD) and Ju Xian Qin Pu, ensuring the inclusion of authentic historical glyphs. Second, to establish ground truth for evaluation, we incorporated scholarly publications that provide parallel translations of these ancient scores into modern staff or numbered musical notation.

Following collection, we manually cropped the documents into snippet-level images to maximize visual clarity and remove irrelevant marginalia. Crucially, to mitigate potential data contamination in Large Multimodal Models (LMMs), our selection process prioritized rare editions and specialized academic resources that are unlikely to be present in standard web-crawled pre-training datasets.

2.2 Terminology of Ancient Chinese Notations

To provide a clear taxonomy for our benchmark, we introduce the five primary notation systems included in **BoyaEval** and detailed examples can be found in Figure 1:

Wenzipu (Textual Notation) *Wenzipu* is the earliest form of *Guqin* notation, dating back to the Tang Dynasty and earlier. It is a "positional system" that uses full Chinese characters to provide a narrative, prose-like description of finger movements and string positions (e.g., "The left thumb presses the seventh string at the tenth fret"). Due to its extreme verbosity, it is often referred to as "Method Notation" (*Shoufapu*).

Jianzipu (Reduced-glyph Notation) Evolved from *Wenzipu*, *Jianzipu* is a revolutionary "instructional algorithm" specifically for the *Guqin*. It utilizes composite ideograms formed by reducing and stacking radicals of Chinese characters. Each glyph encodes four dimensions of information: the finger of the left hand, the fret position, the string number, and the right-hand plucking technique.

Suzipu (Popular Character Notation) Commonly found in the Song and Yuan Dynasties, *Suzipu* is a shorthand, pitch-based system used primarily for folk songs and vocal melodies. It employs ten simplified, cursive-style characters (e.g., *Shao*, *Ye*, *Zha*) to represent specific notes, serving as a precursor to the more standardized *Gongchepu*.

Gongchepu (Gong-Che Notation) Named after two of its core pitch symbols, "*Gong*" and "*Che*", this was the dominant notation system during the Ming and Qing Dynasties. It functions as a character-based solfège system and was widely applied across diverse genres, including traditional opera (*Xiqu*) and instrumental ensembles.

Lulupu (Standard Pitch Notation) Literally meaning "Standard Pitches," *Lülüpu* (Lulupu) is a formal pitch-based system that utilizes the first characters of the Twelve-tone Equal Temperament (the *Shier Lülü*). Because of its rigid and formal nature, it was primarily restricted to court ritual music (*Yayue*) and sacrificial ceremonies rather than secular performance.

2.3 Task Formulation

As shown in Figure 1, BoYaEval is structured around three hierarchical tasks.

Task 1: Structural Parsing (Symbol Recognition) This task evaluates the fundamental visual perception capabilities required to process ancient notation. Unlike modern staff notation, which primarily visualizes pitch and duration on a linear timeline, ancient systems like *Jianzipu* (for Guqin) function as complex logograms. In these systems, a single character spatially compresses multiple pieces of information, such as string numbers, finger positions, and specific plucking techniques.

As shown in the left part of the Figure 2, we formulate this task as a multiple-choice visual question answering (VQA) problem. For data construction, we extracted unique symbols from digitized scores of Guqin and Kunqu opera. Each sample consists of a cropped image of a single notation symbol and a corresponding question asking for its structural decomposition or specific meaning.

To ensure a rigorous evaluation, we employed a hard-negative strategy for distractor generation. Instead of random incorrect options, distractors are constructed by substituting radicals with visually similar components—such as confusing the radical for "Thumb" with "Index Finger"—or using phonetically related characters. This design forces the model to perform fine-grained visual discrimination rather than relying on coarse image features or linguistic priors.

Task 2: Instructional Translation (Semantic Understanding) Moving beyond visual syntax, this task assesses the model's semantic understanding of the operational instructions encoded within the score. As shown in the Figure 2 (middle part), we frame this as an image-to-text generation task where the model must generate a natural language description of the performance technique. We established an expert annotation pipeline involving experts to provide detailed textual descriptions (e.g., "Pluck the seventh string inward with the right middle finger..."). A senior musicology professor reviewed these to verify accuracy. This task tests if MLMMs can bridge the gap between abstract visual representation and concrete musical execution.

Task 3: Musical Reasoning (Melodic Derivation) The final task evaluates the model's ability to synthesize visual cues with domain-specific logic to derive the actual melody. As ancient scores often utilize tablature systems, determining the melody requires reasoning based on instrument tuning and relative intervals. As shown in the right part of Figure 2, we designed multiple-choice questions where the input is an ancient score snippet and the options are segments of modern melody. Distractors include subtle melodic or rhythmic alterations. The model must implicitly "play" the music by applying tuning rules and notation logic to infer the

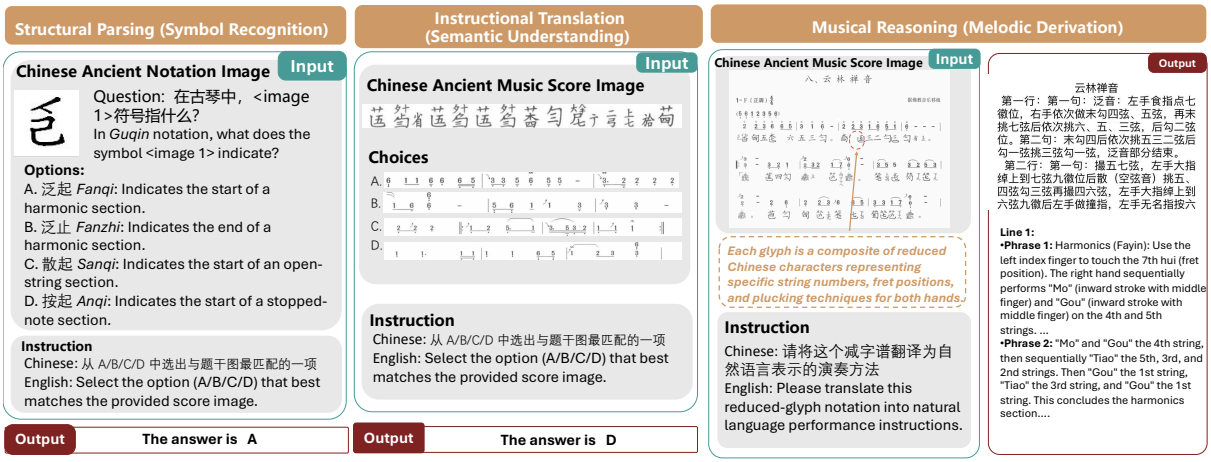

Figure 2: Illustrative examples of the three hierarchical tasks in the **BoyaEval** benchmark.

corresponding modern music notation.

2.4 Dataset Analysis and Statistics

Due to the inherent scarcity of preserved manuscripts and the high barrier to entry for interpretation, constructing a large-scale corpus for this domain is exceptionally challenging. Consequently, we curated a high-quality, expert-annotated dataset comprising 3,175 samples, meticulously sourced from rare manuscripts and authoritative anthologies, as shown in Table 1. The dataset is structured into three distinct tasks designed to evaluate the model's capabilities from fundamental visual perception to high-level musical reasoning.

The first two tasks focus on foundational skills: **Structural Parsing** (503 samples) evaluates basic symbol recognition and historical theory, focusing on the fundamental identification of notation symbols. **Instructional Translation** (1,426 total samples) assesses the model's semantic understanding of operational instructions, requiring it to interpret complex compound ideograms across *Gongchepu*, *Lulupu*, and *Jianzipu*. The final and most challenging component is **Musical Reasoning** (Melodic Derivation), which constitutes the largest portion of the dataset with 1,246 samples (approx. 39%). This task evaluates the model's ability to synthesize visual cues with domain-specific logic to derive the actual melody. We prioritize this category because it represents the ultimate goal of computational musicology: reconstructing sound from "silent history".

Quality over Quantity: Expert Annotation The ground truth was established by three musicology scholars, each with over 10 years of experience. The annotation process involved cross-referencing historical treatises (Qupu) to resolve ambiguities. To ensure a "Gold Standard" reliability, we quantified the inter-annotator agreement using Fleiss' Kappa (κ). Let N be the total number of samples, n be the number of annotators ($n = 3$), and k be the number of categories. The agreement is calculated as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

where $\bar{P}$ is the mean of the extent to which annotators agree for the i-th sample, computed as:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{2}$$

and $\bar{P}_e$ represents the expected agreement by chance:

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \tag{3}$$

In these equations, n_{ij} represents the number of annotators who assigned the i-th sample to the j-th category, and p_j denotes the proportion of all assignments to the j-th category. Our evaluation yielded a κ value of 0.95, indicating almost perfect agreement and validating the dataset's consistency.

Data Characteristics and Challenges Beyond standard character recognition, our corpus preserves intrinsic visual noise such as substrate aging and ink bleed. Crucially, the dataset is characterized by a heavy-tailed distribution of symbol complexity, explicitly curating few-shot exemplars of obscure stylistic variants. This poses a significant challenge for representation learning, requiring models to generalize from sparse data while maintaining invariance to severe visual degradation.

Task 1 Structural Parsing (Symbol Recognition)	Task 2 Instructional Translation (Semantic Understanding)			Task 3 Musical Reasoning (Melodic Derivation)	
Historical Theory Multiple-Choice Questions	Gongchepu Recognition	Lulupu Recognition	Jianzipu Translation	Ancient-Modern Notation Mapping	
				Ancient → Modern	Modern → Ancient
503	943	164	319	1,046	200

Table 1: Detailed statistical breakdown of the **BoyaEval** benchmark across its three-tier hierarchical evaluation framework. The dataset comprises 3,175 total samples meticulously curated from historical manuscripts.

Type	**Model Name**	**Task 1** Knowledge	**Task 2** Recognition		Translation		**Task 3** Reasoning
		Acc.	*BLEU*	*BS*	*BLEU*	*BS*	*Acc.*
Instruct	Moonshot V1 Vision Preview	58.05	2.38	5.77	1.16	19.85	27.29
	Kimi Latest	58.25	4.11	5.97	0.36	19.82	27.29
	Gemini 2.5 Flash	57.65	1.84	7.66	2.14	17.72	29.05
	InternVL2-78B	61.03	7.11	37.87	0.26	24.77	-
	GLM-4.6V	60.04	13.89	42.23	0.97	13.89	28.25
	GPT-4o	49.40	4.22	46.28	0.67	21.90	26.89
	Qwen3-VL-8B-Instruction	56.86	0.14	34.45	0.78	36.77	24.56
	Qwen3-VL-30B-Instruction	61.43	0.07	26.06	0.39	43.76	29.13
	Qwen3-VL-235B-Instruction	63.62	6.45	29.63	5.03	44.29	28.33
	Seed-1.6	**64.40**	15.30	42.63	5.53	40.43	26.08
	Qwen-VL-Max	61.00	7.88	41.80	**9.99**	**49.47**	27.61
	Gemini 2.5 Pro	59.40	**30.48**	**60.57**	4.28	38.45	**29.94**
Thinking	GPT-5	54.40	2.25	3.82	1.48	14.21	20.87
	Gemini 2.5 Flash Thinking	59.24	5.03	6.49	2.70	29.28	26.48
	GLM-4.6V Thinking	63.02	9.08	41.59	0.23	15.48	28.57
	Qwen3-VL-8B-Thinking	60.04	3.84	37.82	1.28	26.98	26.40
	Qwen3-VL-Plus	64.20	7.26	42.42	0.57	26.62	26.32
	Qwen3-VL-235B-Thinking	64.81	5.28	42.22	0.86	29.96	25.60
	Qwen3-VL-30B-Thinking	61.03	2.32	34.05	**9.81**	**49.03**	25.36
	Seed-1.6 Thinking	**65.20**	17.51	46.72	4.07	34.95	27.85
	Gemini 2.5 Pro Thinking	61.40	**30.29**	**61.92**	4.34	37.03	**30.18**

Table 2: Main results on BoYaEval. We report Accuracy (%) for Knowledge and Reasoning tasks. For Recognition and Translation, we report BLEU and BERTScore F1 (BS). The best performance in each category is marked in **bold**, and the second best is underlined. InternVL2 does not support processing more than four images; therefore, it cannot be evaluated on music reasoning tasks in which both the problem statement and each candidate option contain an image.

3 Experiments

3.1 Experimental Setup

Models To comprehensively evaluate the capabilities of current vision-language technologies on ancient music score understanding, we selected a diverse set of Large Multimodal Models (LMMs), which have demonstrated strong performance on OCR and document understanding tasks, specifically Qwen-VL-Max (Bai et al., 2023),Gemini2.5 (Comanici et al., 2025), GLM-4.6V (Team et al., 2025b), Seed-1.6 (Guo et al., 2025), GPT-4o (Hurst et al., 2024), Moonshot, Kimi(Team et al., 2025a) and GPT-5. All models are evaluated in a zero-shot setting to strictly test their inherent knowledge and generalization capabilities.

Evaluation Metrics Given the diverse nature of tasks in our dataset, we adopt a hybrid evaluation strategy consisting of both discriminative and generative metrics:

- **Accuracy for Multiple-Choice Tasks**: For Symbol Recognition and Melodic Derivation, which are formulated as multiple-choice ques-

tions, we report Accuracy (Acc). This metric measures the percentage of correctly selected options (A/B/C/D) against the ground truth.

- **N-gram and Semantic Metrics for Generation Tasks**: For Instructional Translation, which requires the model to translate ancient musical terminology or playing instructions into modern text, we employ standard natural language generation metrics. We report BLEU-4 (Papineni et al., 2002) and ROUGE-L to measure the lexical overlap and structural similarity between the generated explanations and the reference text. Additionally, to capture semantic consistency beyond exact word matching, we report BERTScore (F1), which evaluates the similarity of embedding representations.

Implementation Details For open-source models, inference is conducted on NVIDIA A100 (80GB) GPUs. We set the temperature to 0 to ensure deterministic outputs. The prompts are carefully designed to include task instructions and the image input, formatted consistently across all models. For the Melodic Derivation task, the output is constrained to standard numbered musical notation (1-7) to facilitate automated evaluation.

4 Results

Table 2 presents the comprehensive evaluation results across the three-tier hierarchy of the BoYaEval benchmark. A significant performance gap is observed between foundational Structural Parsing and high-level Musical Reasoning across all evaluated models. For instance, Gemini 2.5 Pro achieves a commendable 59.40% accuracy in symbol recognition (Task 1), demonstrating robust OCR capabilities for non-standard historical glyphs. However, its performance drops precipitously to 29.94% in Musical Reasoning (Task 3). This 29.46% performance gap highlights a critical "reasoning bottleneck": while state-of-the-art LMMs can visually perceive complex symbols, they struggle to ground these visual inputs into the rigorous musicological rules required for melodic derivation. Interestingly, Qwen-VL-Max, an open-source model with strong Chinese language alignment, demonstrates competitive semantic understanding by achieving a 45.64 BERTScore in Instructional Translation (Task 2). This performance rivals or even exceeds proprietary models like GPT-4o in specific translation metrics, suggesting that culturally-aligned pre-training data plays a pivotal role in the semantic interpretation of intangible cultural heritage. Furthermore, the "Thinking" variants of models, such as Seed-1.6-Thinking and Qwen3-VL-Thinking, show incremental improvements in Task 1 knowledge but fail to bridge the reasoning gap in Task 3, with scores remaining largely under 30%.

Model Name	Recognition		Translation	
	BLEU	*BS*	*BLEU*	*BS*
Qwen3-VL-Plus	13.31	56.00	4.18	<u>48.45</u>
Seed-1.6	21.00	**71.04**	**16.37**	**52.26**
Qwen3-VL-Plus (T)	<u>21.11</u>	58.46	4.73	43.14
Seed-1.6 (T)	**32.74**	<u>69.33</u>	<u>8.84</u>	47.82

Table 3: 3-shot Evaluation Results on BoYaEval. (T) denotes the Thinking mode. We report BLEU and BERTScore F1 (BS) for Recognition and Translation tasks. The best performance in each category is marked in **bold**, and the second best is <u>underlined</u>.

To further investigate the models' adaptability, we compared the zero-shot performance in Table 2 with the 3-shot results in Table 3. The introduction of just three examples yielded a substantial performance boost in Recognition tasks; notably, Seed-1.6 (T) saw its BLEU score surge from 17.51% in zero-shot to 32.74% in 3-shot settings. However, the gains in Translation were more conservative, with BERTScore improvements indicating that while few-shot prompts help anchor semantic mapping, the underlying "instructional algorithm" of notations still requires deep domain-specific reasoning that exceeds simple pattern matching.

Further Exploration As illustrated in Figure 3, the evaluation of knowledge task accuracy reveals a significant performance gap across different notation systems, where all models achieve their peak results on Lulupu due to its standardized pitch-based structure while exhibiting much lower proficiency in recognizing the specialized shorthand of Suzipu and the complex character-based system of Gongchepu. Among the evaluated models, Seed-1.6-Think consistently demonstrates the most robust capability across nearly all categories, particularly in formal systems like Lulupu and Wenzipu, suggesting a superior grasp of specialized historical symbolic logic compared to other leading multimodal models.

Figure 4 illustrates the comparative performance of leading Large Multimodal Models (LMMs)

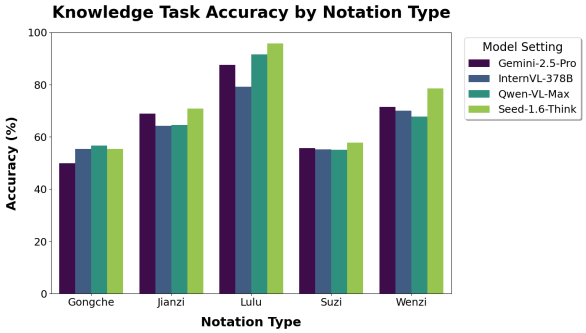

Figure 3: Knowledge task accuracy (%) disaggregated by notation type across four leading MLLMs.

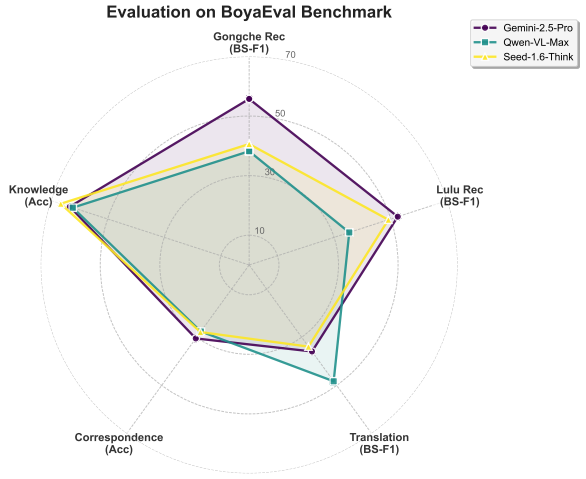

Figure 4: Performance comparison of LMMs on the **BoyaEval** benchmark. The radar chart evaluates models across five key metrics.

across the five evaluation axes of the **BoyaEval** benchmark. The visualization reveals a distinct performance hierarchy among the three hierarchical tasks:(1) Most models, particularly Gemini-2.5-Pro, achieve their highest performance on the Knowledge (Acc) axis. This suggests that state-of-the-art LMMs possess a foundational awareness of ancient Chinese musical symbols and their historical context, likely due to exposure to large-scale cultural corpora during pre-training. (2) Performance across Gongchepu Rec, Lulupu Rec, and Translation (measured via BERTScore) shows that models can reasonably map isolated symbols to semantic instructions. However, as the complexity of the compositional ideograms increases, the models begin to show limitations in capturing the full instructional algorithm encoded in notations like Jianzipu. (3) The most significant bottleneck is observed on the Correspondence (Acc) axis. All evaluated models, including Gemini-2.5-Pro, Qwen-VL-Max, and Seed-1.6, exhibit a sharp performance drop in this area. This failure in melodic derivation confirms that while models can recognize static graphical components, they struggle to apply domain-specific logic to derive the underlying melody from the "silent history" of these scores.

Human vs. Metric Correlation To validate the effectiveness of automatic evaluation metrics in the context of ancient musical score translation, we conducted a granular correlation analysis by randomly sampling 100 data points from the translation task. This analysis compares human expert ratings against the scores generated by the Gemini 2.5 Pro (T) model, which was selected as our primary test vehicle due to its superior comprehensive performance and robust reasoning capabilities.

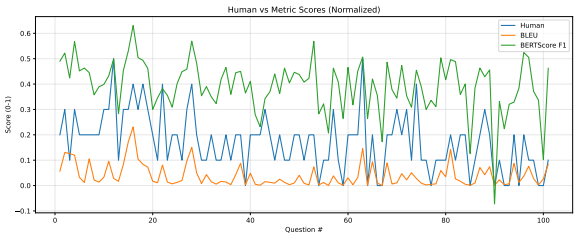

Figure 5: Distribution comparison between human expert scores and automatic metrics (BLEU & BERTScore F1) for the Gemini 2.5 Pro Thinking model on 100 sampled translation instances.

As illustrated in Figure 5, the experimental results reveal distinct behaviors between automated metrics and human judgment. We observe that **BERTScore F1** (green) demonstrates a higher degree of trend alignment with **human scores** (blue) compared to the **BLEU** metric (orange). This phenomenon is primarily attributed to the semantic flexibility required for ancient score translation; while BLEU relies on strict n-gram matching, BERTScore leverages contextual embeddings to capture underlying semantic similarity. The gap between these metrics suggests that while LMMs may not replicate reference translations verbatim, they are highly proficient at capturing the instructional intent encoded in historical symbols. Consequently, the periodic synchronization of peaks between human ratings and BERTScore validates our use of the latter as a primary indicator for semantic understanding, as it more accurately reflects the model's capacity to bridge the cultural gap through cross-modal reasoning.

5 Related Work

Recent Multimodal Large Language Models (MLLMs) are commonly evaluated on broad vi-

sion–language benchmarks that emphasize general perception, reasoning, and instruction following. Representative efforts include MMBench (Liu et al., 2023) and MME (Fu et al., 2023). Beyond generic images, document- and text-centric benchmarks such as DocVQA (Mathew et al., 2021) and OCRBench (Liu et al., 2024) highlight that fine-grained symbol reading (OCR and document VQA) remains a bottleneck even for strong MLLMs. MMMU (Yue et al., 2023) further expands evaluation to diverse expert-domain image types (including music sheets), but its music-related coverage is limited and not designed for systematic music-notation understanding.

Optical Music Recognition focuses on converting score images into machine-readable representations, involving dense small-object recognition and complex structural relations among primitives (e.g., notehead–stem–beam). Large-scale synthetic datasets such as DeepScores (Tuggener et al., 2018) (and its extended DeepScoresV2 (Tuggener et al., 2021)) have been widely used to benchmark symbol detection/segmentation under extreme class imbalance and high symbol density, with DeepScoresV2 adding richer annotations and higher-level information. MUSCIMA++ (Hajič et al., 2018) targets handwritten notation and explicitly annotates symbol relationships (notation graphs), enabling research that bridges low-level primitives and higher-level musical objects as well as graph-based reconstruction. For end-to-end OMR in monophonic settings, PrIMuS (Calvo-Zaragoza and Rizo, 2019) provides printed staff images paired with sequence encodings, while Camera-PrIMuS (Calvo-Zaragoza et al., 2018) introduces realistic camera-like distortions (rotation, blur, illumination changes) to stress robustness.

Music Question Answering (MQA) extends evaluation from transcription to answering structured or semantic queries about music-related inputs. In audio-visual settings, MUSIC-AVQA (Li et al., 2022) benchmarks spatiotemporal reasoning over performance videos with paired audio, requiring models to connect instruments, sounds, and their associations. MusiQAl (Gardner et al., 2021) complements this line by emphasizing music-performance QA that incorporates audio, video, and text, and is designed to probe richer performance-centric understanding beyond purely "what" recognition. For audio-language evaluation, MuChoMusic (Weck et al., 2024) provides human-validated multiple-choice questions over music tracks, aiming to diagnose failure modes such as language-only shortcuts and weak cross-modal integration. More recently, Jamendo-QA (Koh et al., 2025) scales music-audio QA via automatic annotation on openly licensed tracks, supporting both training and zero-shot evaluation for music understanding.

Dedicated benchmarks for visual music notation in the MLLM era are emerging. MusiXQA (Zhang et al., 2025) constructs synthetic music-sheet images with structured annotations (pitch/duration, chords, clefs, key/time signatures, text) and derives diverse visual QA tasks, showing that even strong MLLMs struggle with music-sheet understanding while also enabling targeted fine-tuning. MSU-Bench (Li et al., 2025) evaluates score-level comprehension across both textual (ABC) and visual (PDF) modalities with multi-level generative QA, revealing modality gaps and the difficulty of maintaining correctness across progressive musical-understanding levels. WildScore (Zhao et al., 2025) shifts toward "in-the-wild" score images paired with real user-generated music-theory questions, targeting symbolic music reasoning on realistic inputs. In parallel, broader music-capability evaluations for language models (without focusing on score images) such as ZIQI-Eval (Li et al., 2024b) indicate growing interest in benchmarking music knowledge and reasoning, though these are not tailored to visual notation understanding.

6 Conclusion

In this work, we introduce BoYaEval, a multimodal benchmark designed to digitize and interpret a diverse range of Traditional Chinese Musical notation systems, including Jianzipu, Gongchepu, Wenzipu, Lulupu and Suzipu. By intersecting Large Multimodal Models (LMMs) with intangible cultural heritage, we address the critical challenge of preserving these low-resource historical scripts. Our evaluation uncovers a significant "reasoning gap": while current LMMs demonstrate basic visual recognition capabilities, they exhibit severe deficits in parsing the non-linear spatial structures and decoding the highly compressed semantics of these ancient scores. BoYaEval thus serves as both a rigorous testbed for compositional reasoning and a catalyst for NLP for Social Good. We hope this dataset inspires the development of more robust and culturally inclusive AI systems, ensuring that diverse human legacies are not only preserved but revitalized for future generations.

Limitations

Despite being the first comprehensive benchmark for ancient Chinese musical scores, this study has several limitations that provide avenues for future research. Our benchmark primarily focuses on the transition from visual symbols to semantic text and modern notation (visual-to-symbolic). However, the ultimate goal of ancient music restoration is the auditory realization. **BoYaEval** currently lacks an integrated audio evaluation component (e.g., comparing model-derived melodies against actual expert performances). Future iterations could incorporate audio-visual cross-modal tasks to test if LMMs can directly "hear" the music from the score.

Ethics Statement

The development of the BoYaEval dataset adheres to strict ethical guidelines regarding data sourcing, annotator welfare, and privacy protection. The raw data is primarily derived from historical manuscripts that have entered the public domain, alongside select scholarly materials utilized under the principles of fair use exclusively for non-commercial academic research. We have conducted a rigorous manual review to ensure the dataset is free from personally identifiable information (PII) and contains no offensive or harmful content.

Regarding the annotation process, we engaged domain experts with specialized knowledge in classical Chinese literature and cultural heritage. All annotators were compensated at rates significantly above the local minimum wage and were fully informed about the project's scope and data usage. By releasing BoYaEval, our goal is to support the digital preservation of intangible cultural heritage and advance the capabilities of large language models in understanding complex historical contexts, without infringing on intellectual property rights or compromising individual privacy.

We emphasize that the digitization of these scores is intended to assist, not replace, human inheritance. We oppose the use of these models to generate inauthentic 'fake ancient music' that distorts historical facts.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jorge Calvo-Zaragoza and David Rizo. 2019. Primus: Optical music recognition of monophonic scores. *IEEE Transactions on Multimedia*.

Jorge Calvo-Zaragoza and 1 others. 2018. Camera-primus: Neural end-to-end optical music recognition on realistic monophonic scores. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Geoffrey Martin, and Yifan Peng. 2025. A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends. *arXiv preprint arXiv:2507.09861*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv*.

Matthew Gardner and 1 others. 2021. Musiqal: Music performance question answering. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Jan Hajič and 1 others. 2018. The muscima++ dataset for handwritten optical music recognition. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120.

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

Junyoung Koh and 1 others. 2025. Jamendo-qa: A large-scale music audio question answering dataset. *arXiv*.

Jiajia Li, Lu Yang, Mingni Tang, Chenchong Chenchong, Zuchao Li, Ping Wang, and Hai Zhao. 2024a. The music maestro or the musically challenged, a massive music evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3246–3257.

Jiajia Li and 1 others. 2024b. Ziqi-eval: Evaluating music knowledge and reasoning of large language models. *arXiv*.

Jianwei Li and 1 others. 2022. Music-avqa: Audio-visual question answering over music scenes. In *ACM International Conference on Multimedia (MM)*.

Xinyi Li and 1 others. 2025. Msu-bench: A musical score understanding benchmark. *arXiv*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, and 1 others. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv*.

Yuliang Liu, Zhang Li, Mingxin Huang, and 1 others. 2024. Ocrbench: On the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Pan Lu, Hritik Bansal, Tony Xia, and et al. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640.

Mathew Mathew and 1 others. 2021. Docvqa: A dataset for vqa on document images. *arXiv*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025a. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Lukas Tuggener, Yvan Putra Satyawan, Alexander Pacha, and 1 others. 2021. Deepscoresv2 dataset and benchmark for music object detection. In *25th International Conference on Pattern Recognition (ICPR)*.

Lukas Tuggener and 1 others. 2018. Deepscores – a dataset for segmentation, detection and classification of tiny objects. In *International Conference on Pattern Recognition (ICPR)*.

Benno Weck and 1 others. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. In *Proceedings of ISMIR*.

Xiang Yue, Yuansheng Ni, Kai Zhang, and et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Yixin Zhang and 1 others. 2025. Musixqa: Advancing visual music understanding in multimodal llms. *arXiv*.

Sheng Zhao and 1 others. 2025. Wildscore: In-the-wild music score understanding benchmark. *arXiv*.

A Disparate Performance Across Notation Systems

Table 4 compares model performance across three traditional Chinese music notation types: Gongche, Jianzi, and Lulu. All tasks take as input an image containing a column-level slice of a musical score, but differ in task formulation. The Jianzi task requires interpretative reading of fingering-based symbols, whereas Gongche and Lulu are formulated as symbol recognition tasks. Across models, performance on Jianzi is consistently lower, particularly in BLEU, indicating that interpretative reading poses a greater challenge than direct visual-to-text recognition.

Among the two recognition-based notations, Lulu achieves consistently higher scores than Gongche. This difference can be partly attributed to visual characteristics of the notation systems: Lulu symbols are typically rendered as clear and relatively large textual characters, facilitating recognition, while Gongche symbols are often smaller and visually similar to surrounding musical or structural marks, leading to increased confusion. Comparing Instruct and Thinking settings, explicit reasoning generally improves performance on Gongche and Lulu, but yields limited gains on Jianzi, suggesting that current reasoning mechanisms are insufficient to fully address the semantic inference required by fingering-based notation.

Type	Model	Gongche		Jianzi		Lulu	
		BLEU	BS	BLEU	BS	BLEU	BS
Instruct	Kimi-k2	2.29	1.16	0.36	19.82	15.18	35.19
	Moonshot-v1-vision	0.47	0.79	1.16	19.85	14.00	36.04
	Gemini 2.5 Flash (NT)	0.86	1.36	2.14	17.72	7.46	43.86
	Qwen3-VL-30B-3B (I)	0.04	26.59	0.39	43.76	0.23	23.00
	GLM-4-6V (NT)	12.76	43.07	0.97	13.89	20.42	37.36
	InternVL-3-78B	6.95	38.75	0.26	24.77	8.00	32.89
	GPT-4o	4.49	48.57	0.67	21.90	2.26	29.34
	Qwen3-VL-8B (I)	0.05	33.81	0.78	36.77	0.64	37.99
	Qwen3-VL-235B (I)	5.49	25.12	5.03	44.29	11.94	55.25
	Seed-1.6 (NT)	14.29	42.46	5.53	40.43	21.17	43.64
	Qwen-VL-Max	6.68	42.29	**9.99**	**49.47**	14.81	38.97
	Gemini 2.5 Pro (NT)	**28.46**	**60.83**	4.28	38.45	**42.10**	**59.08**
Thinking	GPT-5 Thinking	1.26	-0.11	1.48	14.21	8.47	28.55
	Gemini 2.5 Flash (T)	1.99	0.27	2.70	29.28	22.48	42.21
	GLM-4-6V (T)	5.89	41.09	0.23	15.48	26.99	44.38
	Qwen3-VL-8B (T)	1.99	35.75	1.28	26.98	14.89	50.15
	Qwen3-VL-30B-3B (T)	1.44	34.71	9.81	**49.03**	7.45	30.20
	Qwen3-VL-235B (T)	1.08	39.36	0.86	29.96	29.36	58.60
	Qwen3-VL-Plus	2.57	38.68	0.57	26.62	33.37	**63.22**
	Seed-1.6 (T)	15.15	44.84	4.07	34.95	30.37	56.95
	Gemini 2.5 Pro (T)	**28.50**	**62.22**	**4.34**	37.03	**40.59**	60.19

Table 4: Performance comparison across different notation types. Models are evaluated on Gongche, Jianzi, and Lulu notations using BLEU and BERTScore (BS).