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

Jiajia Li[1], Weizhi Xue[2], Yao Yao[3], Qiwei Li[2],
Chong Chen[4], Zuchao Li[2]*, Ping Wang[1]*, Hai Zhao[3]
[1]School of Information Management, Wuhan University, Wuhan, China,
[2]School of Artificial Intelligence, Wuhan University, Wuhan, China,
[3]School of Computer Science, Shanghai Jiao Tong University, Shanghai, China,
[4]Shenyang Conservatory of Music, Shenyang, China,
{cantata, weizhi_xue, qw-line, zcli-charlie, wangping}@whu.edu.cn,
{yaoyao27, zhaohai}@sjtu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) 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 21 leading MLLMs. 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 MLLMs in processing diverse spatial-symbolic dependencies, bridging the gap between ancient wisdom and modern AI for digitizing intangible cultural heritage. The BoYaEval benchmark is publicly available at https://huggingface.co/datasets/MYTH-Lab/BoYaEval.

1 Introduction

The advent of Multimodal Large Language Models(MLLMs), 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., 2024; 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 MLLMs 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 Multimodal Large Language 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. While existing Optical Music Recognition (OMR) research has established robust baselines for recognizing individual symbols in Jianzipu, Gongchepu, and Suzipu (Kuremoto et al., 2025; He et al., 2025), these works primarily focus on visual transcription. There remains a lack of a comprehensive framework to evaluate if MLLMs can bridge the gap from visual symbols to musical semantics and melodic derivation. Con-

[1]* Corresponding author.

[2]This work was supported by the National Natural Science Foundation of China (No. 62306216), the National Science and Technology Major Project (No. 2023ZD0121502), the Fundamental Research Funds for the Central Universities (No.2042025kf0090) and by the National Social Science Fundation of China (No. 24&ZD186).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21858–21873
July 2-7, 2026

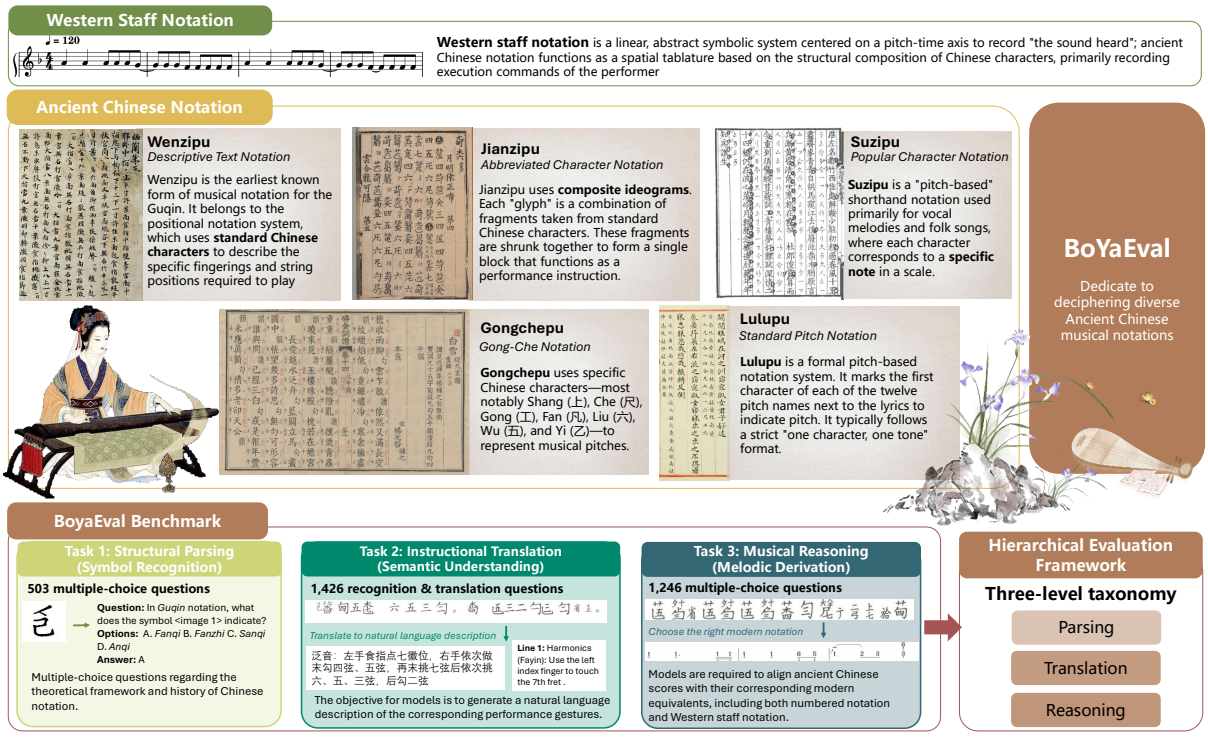

Figure 1: Overview of the BoYaEval benchmark. The hierarchical framework transitions from symbol-level parsing (Task 1) and column-level translation (Task 2) to excerpt-level melodic reasoning (Task 3). This structure reflects the cognitive progression from visual perception to semantic understanding and musicological inference.

sequently, 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) 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 MLLMs (e.g., GPT-4o) are competent in surface-level parsing, they exhibit severe deficits in compositional musical reasoning.

In this work, we make the following contributions:

- Releasing BoYaEval, the first multimodal benchmark for Ancient Chinese Music, containing five types of ancient notations;
- Establishing a three-tier evaluation hierarchy to probe symbolic understanding depth;
- Offering a granular diagnosis of spatial reasoning failures in non-linear ideograms.

2 BoYaEval: Dataset Construction and Analysis

To rigorously evaluate MLLMs 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 Wuhan 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 *Jiugong Dacheng*

Nanbei Ci Gongpu (1746 AD), 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, no additional digital augmentation—such as contrast adjustment, denoising, or sharpening—was applied beyond cropping. This deliberate design choice aims to evaluate MLLMs as end-to-end systems operating on authentic historical document quality. By preserving original visual noise (e.g., substrate aging and ink bleed), we assess the models' intrinsic robustness to realistic archival conditions without introducing potential artifacts or systematic biases from external image restoration pipelines. Besides, to mitigate potential data contamination in MLLMs, 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* (Descriptive Text 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* (Abbreviated Character 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. While individual symbol recognition has been explored in specialized OMR studies—such as Jianzipu recognition using deep learning (Kuremoto et al., 2025) and Gongchepu character identification via optimized YOLO frameworks (He et al., 2025)—BoYaEval integrates these within a broader MLLM-based evaluation.

As shown in the left part of the Figure 2, the task is formulated as a multiple-choice knowledge-based VQA problem. Input: A cropped image of a single notation symbol and a question regarding its structural composition or theoretical meaning. Output: The correct option among four candidates. This task evaluates symbol-level structural knowledge—specifically, the model's ability to disentangle composite ideograms into their constituent radicals (e.g., fingering techniques or string numbers).

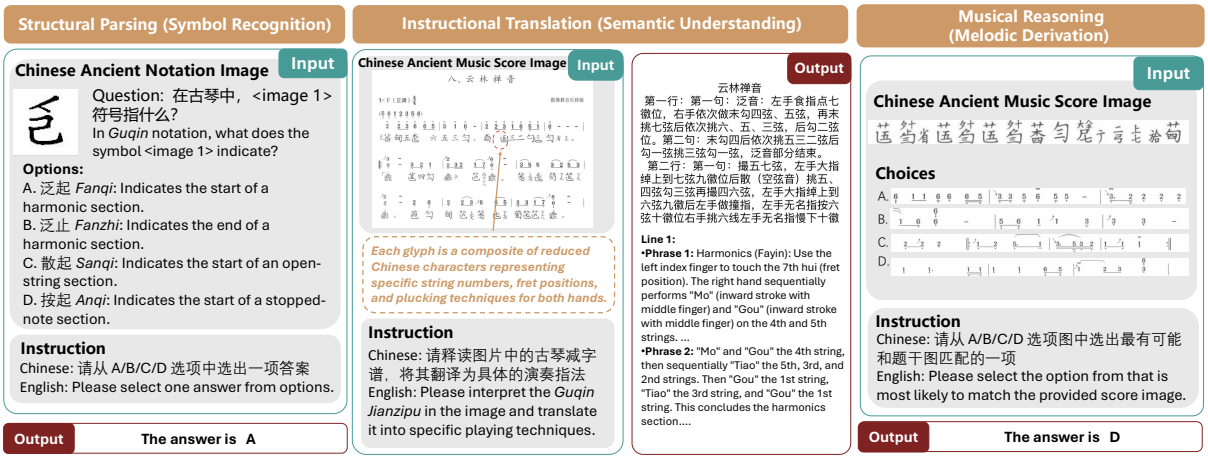

Figure 2: Illustrative examples of the three tasks. (Left) Task 1: VQA on isolated symbols; (Middle) Task 2: Translation of a notation segments or sentences into performance gestures; (Right) Task 3: Alignment between an ancient score excerpt and modern musical notations.

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) As shown in the Figure 2 (middle part), the task focuses on segment or sentence-level image-to-text generation. Input: A cropped image representing a relatively complete musical unit. Output: A natural language description of the performance instructions or an ordered sequence of pitch characters. Unlike the isolated symbols in Task 1, Task 2 assesses semantic decoding of structured musical columns, requiring the model to map spatial-symbolic dependencies to executable physical instructions.

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. Unlike traditional systems that focus on transcribing images into a single digital format (Chen and Sheu, 2014; Repolusk and Veas, 2024, 2025), our Task 3 requires models to reconstruct the underlying melodic structure and align ancient scores with modern staff or numbered notation. 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. Input: A musical excerpt in one system (ancient or modern) and four candidate segments in the target notation. Output: The segment that represents the same underlying melody. This task probes high-level melodic inference, necessitating the synthesis of visual cues with domain-specific musicological logic to derive pitch and rhythm.

Task / Notation System	Samples
Task 1: Structural Parsing	**503**
Jianzipu	231
Suzipu	142
Gongchepu	75
Wenzipu	30
Lulupu	24
Yanyue Banzipu	1
Task 2: Instructional Translation	**1,426**
Gongchepu Recognition	943
Jianzipu Translation	319
Lulupu Recognition	164
Task 3: Musical Reasoning*	**1,246**
Jianzipu ↔ Staff/Numbered	1,041
Gongchepu ↔ Staff	172
Suzipu ↔ Staff	20
Lulupu ↔ Staff	13
Overall Total	**3,175**

* *Note: Task 3 excludes Wenzipu due to lack of aligned transcriptions.*

Table 1: Detailed statistical breakdown of the BoYaEval benchmark. This layout explicitly clarifies the coverage and sample distribution of all notation systems across the three hierarchical tasks.

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. Depending on the specific requirements of Tasks 1–3, the experts performed different roles, including original creation, sequence validation, and alignment curation. Detailed definitions of expert consensus and the step-by-step annotation workflow for each task are provided in Appendix A. 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 recognition, our corpus preserves intrinsic visual noise such as substrate aging and ink bleed. Crucially, the dataset exhibits a heavy-tailed distribution in both symbol frequency and structural complexity. As detailed in Appendix B, while foundational symbols are frequent, a broad tail of complex composite ideograms—sampled from a corpus of over 2,800 Guqin pieces and 5,000 Gongchepu records—appears sparsely. This distribution ensures the benchmark evaluates compositional generalization rather than surface memorization. It requires models to disentangle rare combinations of string, position, and technique, providing a rigorous testbed for developing robust multimodal reasoning techniques under long-tailed historical data constraints.

3 Experiments

3.1 Experimental Setup

Evaluation Paradigm and Models Unlike traditional OMR systems that rely on task-specific architectures (He et al., 2025; Repolusk and Veas, 2025), our benchmark is designed to evaluate the emergent visual-semantic reasoning capacity of general-purpose MLLMs when confronted with diverse, non-linear cultural symbols. We evaluate a diverse set of leading MLLMs (e.g., GPT-4o, Gemini 2.5 Pro, Seed-1.6, and Qwen-VL-Max) primarily in a zero-shot setting to strictly test their inherent generalization capabilities. The complete list of evaluated models, expanded few-shot settings, and detailed implementation setups are provided in Appendix C.

Evaluation Metrics Given the diverse nature of tasks in our dataset, we adopt a hybrid evaluation strategy:

- **Accuracy for Multiple-Choice Tasks**: For Task 1 and Task 3, we report Accuracy (Acc.),

Type	Model Name	Task 1 Knowledge	Task 2 Recognition		Task 2 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
	InternVL3-78B	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. InternVL3 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.

measuring the percentage of correct selections.

- **Metrics for Generation Tasks**: For Task 2 (Instructional Translation), we employ a multi-metric approach: (1) **BLEU-4** (Papineni et al., 2002) for sequential fidelity; (2) **chrF** (Popović, 2015) to capture fine-grained character-level precision of technical terms (e.g., string numbers and fingerings); and (3) **BERTScore (F1)** (Zhang et al., 2019) to evaluate semantic consistency through contextual embeddings.

Design Rationale for Translation Evaluation While structured scoring (e.g., slot-based F1) is conceptually attractive, we deliberately choose the above metrics due to the variable compositional structure of Jianzipu. A single ideogram encodes an inconsistent number of attributes (e.g., finger, string, position, and directional modifiers), making it resistant to a fixed, finite slot schema. Given that our expert-annotated references are formulated as structured performance instructions (e.g., "The left thumb presses the seventh string at the tenth fret"), chrF effectively penalizes precise technical errors, while BERTScore captures the underlying semantic intent. As validated by our correlation study in Section 4, this combination serves as a reliable proxy for musical executability and procedural correctness without the lossy normalization required by an artificial slot-filling framework.

4 Results

Table 2 presents the comprehensive evaluation results. To provide a granular analysis, we disaggregate the three-tier hierarchy into specific subtasks: Task 1 (Knowledge/Parsing), Task 2, which comprises three distinct tracks—Gongchepu Recognition, Lülüpu Recognition, and Jianzipu Translation—and Task 3 (Reasoning/Alignment). Note that while Gongchepu and Lülüpu involve symbol-sequence recognition, Jianzipu Translation evaluates semantic interpretation, leading to

the use of different generative metrics (BLEU vs. BERTScore). 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 MLLMs 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 49.47 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 Adaptability and Headroom To further investigate the models' adaptability, we compared the zero-shot performance with an expanded few-shot study across all task tiers (see Appendix D for complete results). The introduction of just three examples yielded a substantial performance boost in Recognition tasks. 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. Crucially, the most challenging Task 3 (Melodic Derivation) exhibited a distinct reasoning bottleneck, showing marginal or even negative deltas despite the inclusion of 2-shot reasoning traces. This confirms that while lower-level decoding is responsive to adaptation, the high-level cross-representational logic remains the primary performance ceiling.

Further Exploration As illustrated in Figure 3, the evaluation of knowledge task accuracy reveals

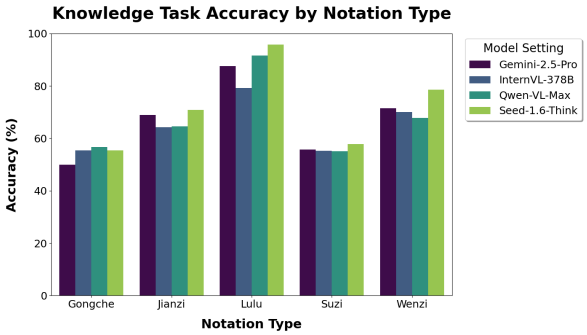

Figure 3: Knowledge task accuracy (%) disaggregated by notation type across four leading MLLMs.

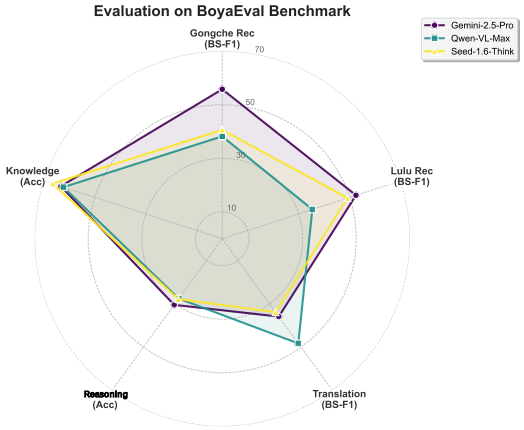

Figure 4: Performance comparison of MLLMs on the BoYaEval benchmark. BERTScore F1, Acc. represents Accuracy (%).

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 MLLMs 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 MLLMs 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 Translation (*Jianzipu*). A detailed comparative analysis of this disparity in generative tasks—specifically contrasting the interpretative reading required for *Jianzipu* against the direct visual recognition of *Gongchepu* and *Lulupu*—is provided in Appendix F.(3) The most significant bottleneck is observed on the Reasoning (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 our metrics, we sampled 100 translation outputs evaluated by domain experts. Though constrained by the extreme scarcity of ancient music experts, this sample provides sufficient statistical power for robust external validation. We evaluate outputs from both Gemini 2.5 Pro and Seed-1.6-Think to ensure our findings are not architecture-dependent. Crucially, this study confirms that our automated framework reliably tracks the procedural correctness of generated music.

As illustrated in Figure 5, BERTScore F1 (orange) demonstrates a higher monotonic alignment with human scores (green) than BLEU (blue) across both models. This gap is attributed to the semantic flexibility required for ancient score translation: while BLEU relies on strict n-gram matching, BERTScore leverages contextual embeddings. This suggests that although MLLMs may not replicate reference translations verbatim, they are highly proficient at capturing the underlying instructional intent encoded in historical symbols. Consequently, the periodic synchronization of peaks between human ratings and BERTScore visually validates its capacity to evaluate cross-modal reasoning and bridge the cultural gap. To provide a rigorous statistical basis for this observation and distinguish rank-order consistency from absolute value accuracy, we

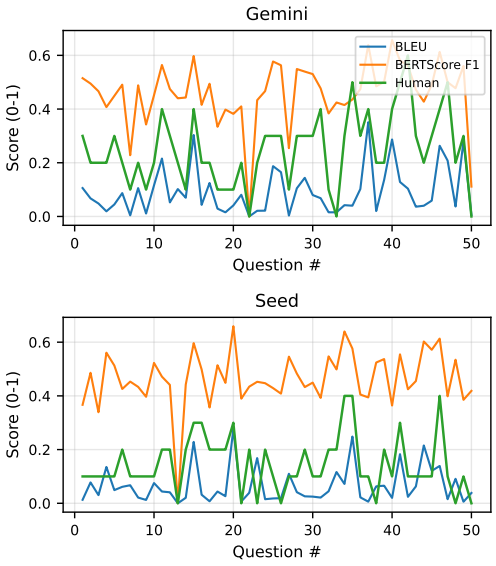

Figure 5: Distribution comparison between human expert scores and automatic metrics (BLEU & BERTScore F1) for the Gemini 2.5 Pro Thinking model on 100 sampled translation instances.

Model	Pearson Correlation			Spearman Correlation		
	BLEU	chrF	**BS**	BLEU	chrF	**BS**
Gemini 2.5 Pro	0.589	0.713	**0.721**	0.690	0.625	**0.770**
Seed 1.6 Think	0.448	0.477	**0.510**	0.344	0.384	**0.451**

Table 3: Correlation coefficients between human expert ratings and automatic metrics (BLEU, chrF, and BERTScore F1 [BS]) for Task 2 translation outputs. BS consistently shows stronger monotonic alignment with human judgments.

calculate Pearson and Spearman correlations.

As shown in Table 3, BERTScore F1 consistently outperforms BLEU across both metrics. We emphasize Spearman's rank correlation to evaluate monotonic alignment rather than absolute value accuracy (e.g., MSE). For Gemini 2.5 Pro, BERTScore achieves a strong Spearman correlation of 0.770 (vs. BLEU's 0.690). Crucially, because our expert rubric explicitly evaluated whether translations were musically executable and structurally correct, this high correlation indicates that BERTScore does not merely track lexical overlap, but effectively captures the procedural correctness of the music. By surpassing BLEU, BERTScore proves that capturing the underlying semantic intent of a fingering gesture is a more reliable proxy for performance-level accuracy than exact verbatim matching.

Mechanistic Failure Analysis To move beyond descriptive reporting and investigate why models fail, we conducted a manual audit of reason-

ing traces from 'Thinking' models (e.g., Seed-1.6-Think and Gemini 2.5 Pro) on Task 3. We categorized recurrent failure modes into structural shortcuts, such as single-cue bias and canonical template override. A comprehensive breakdown of these mechanistic failures and representative case studies are provided in Appendix G.

5 Related Work

5.1 MLLM and Document Benchmarks

Recent Multimodal Large Language Models (MLLMs) are commonly evaluated on broad vision-language benchmarks for general perception and reasoning, such as MMBench (Liu et al., 2024a) and MME (Fu et al., 2025). Beyond generic images, document-centric benchmarks like DocVQA (Mathew et al., 2021) and OCR-Bench (Liu et al., 2024b) highlight that fine-grained symbol reading remains a bottleneck. While MMMU (Yue et al., 2024) expands evaluation to expert domains including music sheets, its coverage is limited and not designed for systematic music-notation understanding. Existing benchmarks predominantly feature modern, print-standard scores. They overlook the intersection of fluid calligraphy and rigid musical grammar found in ancient manuscripts, leaving these specific challenges entirely unaddressed in current evaluation standards.

5.2 Optical Music Recognition (OMR)

OMR focuses on converting score images into machine-readable representations, involving complex structural relations. Large-scale synthetic datasets like DeepScores (Tuggener et al., 2018, 2021) benchmark symbol detection under high density. MUSCIMA++ (Hajič and Pecina, 2017) targets handwritten notation graphs, while PrIMuS and Camera-PrIMuS (Calvo-Zaragoza and Rizo, 2018b,a) address end-to-end monophonic OMR with realistic distortions.

Recently, traditional Chinese OMR has advanced significantly. For Jianzipu, Kuremoto et al. (2025) demonstrated the efficacy of deep learning methods for character recognition. In the domain of Gongchepu, early systems utilized specialized rule-based recognition (Chen and Sheu, 2014), while recent works introduced advanced object detection frameworks like YOLOv8m combined with attention mechanisms to achieve precise character identification (He et al., 2025). Regarding Suzipu, the KuiSCIMA datasets (Repolusk and Veas, 2024, 2025) provided a critical foundation for recognizing historical notations, addressing cross-notation generalization and baseline calibration.

5.3 Music QA and Visual Score Understanding

Music Question Answering (MQA) evaluates structured or semantic queries about music. This includes audio-visual spatiotemporal reasoning (Li et al., 2022; Christodoulou et al., 2025) and audio-language evaluation for tracks (Weck et al., 2024; Koh et al., 2025).

Dedicated benchmarks for visual music notation are also emerging. MusiXQA (Chen et al., 2025) and MSU-Bench (Dai et al., 2025) evaluate score-level comprehension via synthetic images and multi-level generative QA, respectively, showing that strong MLLMs struggle with music-sheet understanding. WildScore (Mundada et al., 2025) shifts to "in-the-wild" score images with user-generated questions. Concurrently, broader music-capability evaluations like ZIQI-Eval (Li et al., 2024) benchmark knowledge but are not tailored to visual notation. BoYaEval fills this gap by focusing explicitly on the visual-semantic reasoning and structural parsing of diverse ancient Chinese musical scores.

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 Multimodal Large Language Models 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 MLLMs 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 MLLMs 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 experts were compensated at a professional rate of approximately 300 RMB per hour. Beyond compensation, all participants were fully informed of the research intent and provided written informed consent for the use and public release of their annotations in the context of evaluating MLLMs. 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*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Jorge Calvo-Zaragoza and David Rizo. 2018a. Camera-primus: Neural end-to-end optical music recognition on realistic monophonic scores. In *ISMIR*, pages 248–255.

Jorge Calvo-Zaragoza and David Rizo. 2018b. End-to-end neural optical music recognition of monophonic scores. *Applied Sciences*, 8(4):606.

Gen-Fang Chen and Jia-Shing Sheu. 2014. An optical music recognition system for traditional chinese kunqu opera scores written in gong-che notation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):7.

Jian Chen, Wenye Ma, Penghang Liu, Wei Wang, Tengwei Song, Ming Li, Chenguang Wang, Jiayu Qin, Ruiyi Zhang, and Changyou Chen. 2025. Musixqa: Advancing visual music understanding in multimodal large language models. *arXiv preprint arXiv:2506.23009*.

Anna-Maria Christodoulou, Kyrre Glette, Olivier Lartillot, and Alexander Refsum Jensenius. 2025. Musiqal: A dataset for music question–answering through audio–video fusion. *Transactions of the International Society for Music Information Retrieval*, 8(1).

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*.

Congren Dai, Yue Yang, Krinos Li, Huichi Zhou, Shijie Liang, Zhang Bo, Enyang Liu, Ge Jin, Hongran An, Haosen Zhang, and 1 others. 2025. Musical score understanding benchmark: Evaluating large language models' comprehension of complete musical scores. *arXiv preprint arXiv:2511.20697*.

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, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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 Pavel Pecina. 2017. The muscima++ dataset for handwritten optical music recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 39–46. IEEE.

Zhizhou He, Yuqian Zhang, Liumei Zhang, and Yuanjiao Hu. 2025. Precise recognition of gong-che score characters based on deep learning: Joint optimization of yolov8m and simam/mscam. *Electronics*, 14(14):2802.

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, Soo Yong Kim, Yongwon Choi, and Gyu Hyeong Choi. 2025. Jamendo-qa: A large-scale music question answering dataset. *arXiv preprint arXiv:2509.15662*.

Takashi Kuremoto, Kazuma Fujino, Hirokazu Takahashi, Shun Kuremoto, Mamiko Koshiba, Hiroo Hieda, and Shingo Mabu. 2025. Recognition of guqin music notation of jianzi pu by deep learning methods. *Journal of Robotics, Networking and Artificial Life*, 11(1):83–88.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19108–19118.

Jiajia Li, Lu Yang, Mingni Tang, Chenchong Chenchong, Zuchao Li, Ping Wang, and Hai Zhao. 2024. 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.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024a. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 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.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Gagan Mundada, Yash Vishe, Amit Namburi, Xin Xu, Zachary Novack, Julian McAuley, and Junda Wu. 2025. Wildscore: Benchmarking mllms in-the-wild symbolic music reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16858–16874.

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.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Tristan Repolusk and Eduardo Veas. 2024. The kuiscima dataset for optical music recognition of ancient chinese suzipu notation. In *International Conference on Document Analysis and Recognition*, pages 38–54. Springer.

Tristan Repolusk and Eduardo Veas. 2025. Kuiscima v2. 0: Improved baselines, calibration, and cross-notation generalization for historical chinese music notations in jiang kui's baishidaoren gequ. In *International Conference on Document Analysis and Recognition*, pages 116–132. Springer.

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, Ismail Elezi, Jurgen Schmidhuber, Marcello Pelillo, and Thilo Stadelmann. 2018. Deepscores-a dataset for segmentation, detection and classification of tiny objects. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3704–3709. IEEE.

Lukas Tuggener, Yvan Putra Satyawan, Alexander Pacha, Jürgen Schmidhuber, and Thilo Stadelmann. 2021. The deepscoresv2 dataset and benchmark for music object detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9188–9195. IEEE.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2408.01337*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Expert Annotation Workflow

To ensure the high reliability of the BoYaEval benchmark, we employed a rigorous expert annotation pipeline. The construction of the ground truth (GT) for each task followed distinct procedural logics:

A.1 Task 1: Expert-Created Ground Truth

All Task 1 questions were **fully constructed** by domain experts. This process involved: (1) Designing specialized multiple-choice questions centered on historical theory; (2) Crafting theoretically plausible but incorrect distractors based on common notation ambiguities; and (3) Manually determining the correct answer. The GT for Task 1 is therefore a purely expert-authored dataset.

A.2 Task 2: Consensus-Based and Expert-Validated GT

Task 2 utilized two different workflows based on the subtask type:

- **Translation (e.g., Jianzipu):** Experts performed manual interpretation of cropped notation segments. For these textual explanations, "annotator agreement" refers to a **consensus-based validation**. Two experts independently reviewed each interpretation; any discrepancies in fingering descriptions or technique mapping were discussed until a finalized consensus version was reached as the GT.

- **Recognition (e.g., Gongchepu):** An initial sequence was generated via an MLLM (Gemini 3 Pro) and subsequently **expert-validated**. Experts reviewed every predicted symbol, corrected errors in sequence and structural integrity, and verified the final character mapping.

A.3 Task 3: Expert-Curated Alignment

Task 3 items were **automatically constructed** from paired ancient and modern score scans and subsequently **expert-curated**. Experts systematically reviewed these generated alignment questions to: (1) Verify that the ancient and modern excerpts correspond to the same underlying melody; (2) Ensure the correct option is musically unambiguous; and (3) Refine or remove any items with rhythmic or melodic ambiguity.

B Detailed Corpus Statistics and Distribution

To substantiate the heavy-tailed nature of the BoYaEval benchmark, we provide additional statistics regarding our source corpus and symbol distribution.

B.1 Historical Coverage and Diversity

The benchmark is sampled from a vast historical repository, ensuring wide stylistic and structural diversity:

- **Jianzipu**: 322 representative excerpts selected from over 2,800 historical pieces, spanning from the Ming dynasty (e.g., *Shenqi Mipu*) to the late Qing period.

- **Gongchepu**: 946 excerpts curated from 5,000+ pieces, including major Qing-era compilations like *Jiugong Dacheng Nanbei Ci Gongpu*.

This coverage spans over 1,300 years, capturing the evolutionary trajectory of musical notation.

B.2 The Heavy-Tailed Characteristic

The "heavy-tailed" property manifests in two dimensions:

1. **Symbol Frequency Tail**: A small subset of foundational symbols (e.g., basic string markers) dominates the high-frequency spectrum, while a broad tail of structurally rich, historically specific composite symbols appears sparsely.

2. **Structural Complexity Skew**: Low-frequency symbols often possess the highest structural density, integrating multiple components (string, *hui* position, left/right-hand techniques) into a single ideogram.

This dual skew ensures that the benchmark evaluates a model's ability to generalize across rare, complex symbolic combinations rather than relying on surface patterns or frequency biases.

C Detailed Experimental Setup

This section provides comprehensive details regarding the evaluated models and implementation configurations, expanding upon the summary in Section 3.1.

C.1 Evaluated Models and Settings

To comprehensively evaluate the capabilities of current vision-language technologies on ancient music score understanding, we selected a diverse set of MLLMs, which have demonstrated strong performance on OCR and document understanding tasks. Specifically, our evaluation includes InternVL3 (Zhu et al., 2025), Qwen-VL-Max (Bai et al., 2023), Qwen3-VL series (8B, 30B, 235B and Plus) (Bai et al., 2025), Gemini 2.5 Pro and Flash (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 initially evaluated in a zero-shot setting to strictly test their inherent knowledge. To further investigate performance headroom and adaptation capacity, we conduct an expanded few-shot analysis. Specifically, we evaluate Task 2 under a 3-shot setting and Task 3 under a 2-shot setting, providing models with fully worked melodic reconstruction and reasoning traces as demonstrations. The complete model coverage and detailed analysis for these few-shot settings are provided in **Appendix D**.

C.2 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. Critically, for Task 2 and Task 3, the prompts explicitly inform the model of the specific notation system being processed to provide necessary domain context. For Task 1, the notation context is inherent within the theoretical question itself. Detailed prompt templates and examples for each task and notation type are provided in **Appendix E**. For the Melodic Derivation task (Task 3), the output is constrained to standard numbered musical notation (1-7) to facilitate automated evaluation.

D Systematic Few-shot Study

To provide a comprehensive view of model adaptation capacity beyond zero-shot performance, we expanded our evaluation to include a systematic few-shot study across all task tiers. Task 2 is evaluated in a **3-shot** setting, while Task 3 utilizes a **2-shot** setting with explicit reasoning chains.

Model	Jianzipu (Trans.)		Gongchepu (Rec.)		Lulupu (Rec.)	
	BLEU	BS	BLEU	BS	BLEU	BS
Gemini 2.5 Pro	7.94	45.63	53.05	80.30	45.34	79.51
Gemini 2.5 Pro (T)	7.95	45.36	53.15	80.19	46.72	79.68
Seed-1.6	5.12	41.20	48.90	76.45	40.12	72.34
Seed-1.6 (T)	5.45	42.15	49.34	77.10	42.56	74.12
GLM-4.6V	8.92	46.64	30.48	64.30	24.17	75.12
GLM-4.6V (T)	1.75	32.85	27.40	62.56	54.23	82.63

Table 4: Complete 3-shot results for Task 2. **BS** stands for BERTScore F1.

D.1 Task 2: Detailed Performance Headroom

Table 4 presents the complete results for 6 model configurations. We observe that structured recognition tasks (Gongchepu and Lulupu) exhibit the highest "adaptation headroom," with several models showing improvements of over +20 points in BLEU after seeing only three examples.

Model	Accuracy (%)	Delta (Δ)
Gemini 2.5 Pro (T)	30.87	+0.69
Seed-1.6	30.08	+4.00
Seed-1.6 (T)	27.13	-0.72
GLM-4.6V	26.79	-1.46
GLM-4.6V (T)	27.24	-1.33
Qwen3-VL Plus (T)	25.99	-0.33

Table 5: Task 3 Few-shot (2-shot) Accuracy. Δ indicates changes relative to zero-shot.

D.2 Task 3: Reason-Informed Few-shot Analysis

For the highest tier, we provide models with two demonstrations that include fully worked melodic reconstruction and interval reasoning traces. As shown in Table 5, despite the inclusion of structured reasoning steps, performance gains are marginal or even negative. This suggests a significant "reasoning bottleneck" where models fail to internalize the cross-representational logic of melodic derivation from sparse examples.

E Prompt Templates and Examples

We provide the full set of prompting strategies and templates used across the three hierarchical tasks to ensure the transparency and reproducibility of our experimental setup.

E.1 Task 1: Structural Parsing

Task 1 is framed as a knowledge-based multiple-choice VQA problem. The notation context is included within the question text.

Template:
Please select one answer from options A/B/C/D based on the provided image.
Output one sentence explanation and the final answer in the following format:
Reason: <your explanation>
Answer: <option letter>

E.2 Task 2: Instructional Translation

The prompt explicitly specifies the notation system and the expected output granularity (performance gestures or pitch symbols).

Jianzipu Translation Template:
Please interpret the Jianzipu (reduced-character notation) shown in the image and translate it into concrete performance fingering instructions.
Output format: Interpretation: <translation result>

Gongchepu/Lülüpu Recognition Template:
Please recognize the [notation type] column in the image and output the notation symbols in sequential order.
Output format: Notation: <recognized symbols>

E.3 Task 3: Musical Reasoning

Task 3 specifies both the source notation in the image and the target notation in the options to guide the cross-notation alignment.

Template:
The image shows [notation type 1]. Please select the most likely corresponding [notation type 2] from the following options.
Output format:
Reason: <your explanation>
Answer: <option letter>

F Disparate Performance Across Notation Systems

Table 6 compares model performance across three traditional Chinese music notation types: *Gongchepu*, *Jianzipu*, and *Lulupu*. All tasks take as input an image containing a column-level slice of a musical score, but differ in task formulation. The *Jianzipu* task requires interpretative reading of fingering-based symbols, whereas *Gongchepu* and *Lulupu* are formulated as symbol recognition tasks. Across models, performance on *Jianzipu* 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, *Lulupu* achieves consistently higher scores than *Gongchepu*. This difference can be partly attributed to visual characteristics of the notation systems: *Lulupu* symbols are typically rendered as clear and relatively large textual characters, facilitating recognition, while *Gongchepu* 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 *Gongchepu* and *Lulupu*, but yields limited gains on *Jianzipu*, suggesting that current reasoning mechanisms are insufficient to fully address the semantic inference required by fingering-based notation.

G Mechanistic Failure Analysis of Task 3

To investigate the underlying causes of performance bottlenecks in Task 3 (Melodic Reasoning), we analyzed the explicit reasoning traces of "Thinking" models. By examining the intermediate inference steps of 20 representative failure cases each for Seed-1.6-Think and Gemini 2.5 Pro, we identified four primary failure mechanisms:

G.1 Single-Cue Bias and Premature Commitment

A dominant failure mode (observed in 8/20 cases for Seed-1.6-Think) is the disproportionate weighting of salient visual symbols. For instance, the presence of a *harmonic circle* often triggers an immediate assumption of a harmonic segment, leading the model to ignore contradictory pitch structures. Similarly, ornaments like slides ("↑") act as anchors that cause the model to commit to a melodic contour before verifying the complete symbol-to-pitch mapping chain.

G.2 Top-Down Template Override

Especially prevalent in Gemini 2.5 Pro, models often exhibit **canonical repertoire bias**. Instead of performing bottom-up decoding of the provided score, the model maps the excerpt to well-known pieces (e.g., *Flowing Water* or *Yangguan Sandie*) stored in its training data. This top-down prior frequently overrides local notation evidence, leading to post-hoc rationalizations when the actual notation contradicts the memorized template.

Type	Model	Gongchepu		Jianzipu		Lulupu	
		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	0.86	1.36	2.14	17.72	7.46	43.86
	Qwen3-VL-30B-A3B	0.04	26.59	0.39	43.76	0.23	23.00
	GLM-4.6V	12.76	43.07	0.97	13.89	20.42	37.36
	InternVL3-78B	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	0.05	33.81	0.78	36.77	0.64	37.99
	Qwen3-VL-235B	5.49	25.12	5.03	44.29	11.94	55.25
	Seed-1.6	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	**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	1.99	0.27	2.70	29.28	22.48	42.21
	GLM-4.6V	5.89	41.09	0.23	15.48	26.99	44.38
	Qwen3-VL-8B	1.99	35.75	1.28	26.98	14.89	50.15
	Qwen3-VL-30B-3B	1.44	34.71	9.81	**49.03**	7.45	30.20
	Qwen3-VL-235B	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	15.15	44.84	4.07	34.95	30.37	56.95
	Gemini 2.5 Pro	**28.50**	**62.22**	**4.34**	37.03	**40.59**	60.19

Table 6: Performance comparison across different notation types. Models are evaluated on Gongche, Jianzi, and Lulu notations using BLEU and BERTScore (BS).

G.3 Unstable Foundation and Evidence-Chain Breakdown

Failure often stems from a fragile symbolic foundation. Reasoning traces frequently contain uncertainty markers (e.g., "maybe," "reconsider") during the initial decomposition of *Jianzipu*. This instability propagates downstream, where the model eventually substitutes rigorous interval-consistent alignment with a holistic, qualitative "contour matching" approach (e.g., "this option has a similar flow").

G.4 Inconsistent Tuning and Pitch Mapping

Models struggle to maintain internal consistency regarding instrument tuning and the *hui*-to-pitch correspondence. This leads to incorrect starting tones or unstable melodic skeletons, where the model acknowledges a pitch mismatch but fails to recalibrate its reasoning, often invoking "non-standard tuning" as an ad-hoc justification for an incorrect final answer.