computers

MDPI

Article

A Multi-Source Pipeline for Extracting Traditional-Style Chinese Melody Data from Symbolic Files and Score Images

Xuanfei Zhou [1], Yinxuan Huang [2], Sining Han [3] and Jiangyao Bai [3,*]

[1] Preschool College, Changsha Normal University, Changsha 410100, China
[2] College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China
[3] College of Systems Engineering, National University of Defense Technology, Changsha 410073, China
* Correspondence: baijiangyao@nudt.edu.cn

Abstract

Large-scale symbolic melody datasets are essential for data-driven music information retrieval and generation, yet traditional-style Chinese melodies remain scattered across heterogeneous score formats and image sources. Existing extraction pipelines typically focus on single modalities—either MIDI archives or standard staff notation—and lack unified handling for numbered musical notation (Jianpu) and automated quality assurance. We propose the Multi-Source Melody Pipeline (MSMP), a *systems-integration prototype* whose front-end admits MIDI, MusicXML, Jianpu images, and staff images, and whose back-end converges on a standardized event-level representation; the present case study exercises the image branch—in particular the Jianpu branch, through a Gemini-2.5-flash vision language model—and treats the MIDI/MusicXML ingestion paths as architectural slots that are wired in but not experimentally validated in this submission. The system employs notation-aware routing to direct score images to appropriate backends (a VLM for Jianpu and rule-based OMR for staff) and enforces a *structural* validity gate (schema conformance plus at least one melodic track with at least one musical event) on every candidate segment. Validation on a 292-page representative prototype cohort yielded an 80.1% *structural-acceptance* rate—explicitly not a transcription accuracy number—and a newly added ground-truth benchmark on 50 manually annotated Jianpu pages reports 95.8% time-signature exact accuracy, 77.1% tonal-pitch-class key accuracy, 100% tempo agreement within ± 5 BPM, and, on a 10-page note-level subset, a mean first-16-note pitch F1 of 0.898 (octave-sensitive) with a Symbol Error Rate of 0.150. A companion 10-page $K = 3$ self-consistency audit indicates that metadata errors are systematic rather than stochastic. This work, therefore, contributes a reproducible integration *architecture* and a quantitative baseline on the Jianpu branch, rather than a new OMR algorithm, a new dataset release, or a fully benchmarked multi-format corpus; ongoing work addresses out-of-distribution classifier evaluation, comparison against dedicated Jianpu OMR baselines, and release of a copyright-cleared corpus.

Keywords: melody extraction; dataset construction method; optical music recognition; traditional-style Chinese music; Jianpu transcription; automated pipeline

Check for updates

Academic Editor: Paolo Bellavista

Received: 3 April 2026
Revised: 24 April 2026
Accepted: 24 April 2026
Published: 7 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid advancement of music information retrieval (MIR) and generative Artificial Intelligence (AI) has heightened the demand for large-scale, high-quality melodic

datasets [1–3]. Although classical treatments of MIR and music processing [4,5] still anchor many methodological choices, the field has since moved toward large symbolic corpora, pre-trained audio–text models, and generative sequence models that rely on well-curated note-level data [3,6]. Such corpora underpin tasks ranging from automatic transcription and melody completion to style-conditioned generation [6,7]. While Western classical and popular music benefit from numerous openly available datasets—including MIDI collections such as the Lakh MIDI Dataset (176,581 files) [8] and curated piano corpora like MAESTRO [9]—traditional-style Chinese music remains severely under-represented. This scarcity hinders the development of culturally inclusive AI models capable of understanding and synthesizing non-Western musical idioms.

The traditional-style Chinese music repertoire encompasses a broad range of compositions drawing on pentatonic and heptatonic scales, characteristic rhythmic patterns, and tonal inflections absent from mainstream Western harmony [10,11]. Its written sources span multiple notation systems simultaneously in circulation: standard staff notation, Jianpu (numbered musical notation), digital MIDI archives, and MusicXML files exported from engraving software. Each format presents distinct parsing challenges. Staff images are relatively well-supported by existing Optical Music Recognition (OMR) systems [12]; however, Jianpu images—which encode pitch as Arabic numerals 1–7, modify octave through a system of supra- and infra-linear dots, and represent duration via underlines and trailing dashes—remain comparatively under-served by specialized OMR approaches [13]. To the best of our knowledge, no prior work has unified all four input modalities into a single automated pipeline with integrated downstream quality assurance.

The absence of such a pipeline has practical consequences. Researchers studying traditional Chinese music must either manually digitize scores—a process that is slow, error-prone, and unscalable—or limit themselves to the small number of existing symbolic collections for traditional Chinese music [14]. This challenge is exacerbated by the fact that Jianpu remains the predominant notation format in popular Chinese music publishing, particularly for folk songs, film scores, and educational materials. Without reliable Jianpu OMR, a vast cultural archive remains locked in paper form.

In this study, we propose an end-to-end data extraction framework called the *Multi-Source Melody Pipeline* (MSMP). MSMP accepts MIDI files, MusicXML documents, and raw sheet music images (both Staff and Jianpu) as inputs, and produces a validated, schema-consistent melody corpus. The pipeline is organized into three stages. First, symbolic files (MIDI and MusicXML) are parsed and normalized via the `music21` library [15]. Second, score images are classified by a zero-shot vision language model (VLM) and routed to format-specific OMR backends: Jianpu images undergo VLM-based direct transcription, while Staff images are processed by the rule-based Audiveris engine [16]. Third, a melody-oriented quality control (QC) module verifies structural validity before all branches converge on a unified event-level representation.

A prototype case study further shows that MSMP can recover structurally coherent melody segments from raw score images and generate interpretable descriptive statistics from the extracted outputs. Accordingly, the emphasis of this paper is on the *systems-integration methodology* for heterogeneous melody extraction and normalization, together with a quantitative accuracy baseline on the Jianpu branch, rather than on proposing a new OMR algorithm, a new musicological model, or a finished publicly released corpus.

We scope the contributions of this work deliberately narrowly, and state up-front what this paper is *not*. This is a *systems-integration* contribution: it is not a new Jianpu OMR architecture, it is not a new symbolic-music benchmark, and it is not a comparative musicological study. The primary contributions, under this explicit scope, are:

1. A **multi-source ingestion architecture** whose symbolic branch (MIDI, MusicXML) and image branch (Jianpu, Staff) produce a single event-level schema. The image branch is experimentally validated in this paper; the MIDI/MusicXML branch is wired in but not separately benchmarked in the present submission and is, therefore, reported as an architectural slot rather than as quantitatively validated evidence.
2. A **VLM-driven Jianpu transcription strategy** that uses a general-purpose vision language model to decode numbered notation without handcrafted symbol segmentation or task-specific fine-tuning, together with an explicit content-level accuracy measurement against a manually annotated ground-truth subset and an alt-prompt ablation that isolates how much of that accuracy is attributable to the production prompt.
3. A **unified event representation and a structural QC layer** that normalize heterogeneous outputs to a single downstream schema. The QC layer is explicitly *structural*—schema plus at least one melodic track with at least one musical event—rather than a content-level correctness gate; content-level correctness is reported separately in Section 4.7.
4. A **quantitative accuracy and consistency audit of the Jianpu branch**, comprising a 50-page manually annotated ground-truth benchmark (key/time-signature/BPM/first-16-note pitch F1 and SER; Section 4.7) and a 10-page $K = 3$ self-consistency audit (Table 4), so that the reliability of the Jianpu branch is measured against an external reference rather than inferred from the pipeline's own acceptance rate.

We correspondingly do *not* claim in this paper that MSMP outperforms dedicated Jianpu OMR systems, that it produces publicly releasable data, or that it has been tested end-to-end on out-of-distribution score types (balanced staff/mixed/invalid test pages). These are declared and scoped as future work in Section 6.

The remainder of this paper is organized as follows. Section 2 reviews related work on symbolic music datasets, OMR, and automated dataset construction. Section 3 details the system architecture, OMR methodologies, and QC criteria. Section 4 presents comprehensive dataset statistics and musicological analyses. Section 5 situates the findings within the broader research landscape and discusses limitations. Section 6 concludes the paper.

2. Related Work

2.1. Symbolic Music Datasets

Publicly available symbolic music datasets span a much broader stylistic range than the classical/popular axis alone, and can be grouped by notation type, cultural provenance, and target task. Large MIDI repositories—such as the Lakh MIDI Dataset (176,581 files) [8], MAESTRO (about 200 h of aligned piano audio and MIDI) [9], and GiantMIDI-Piano (10,854 classical piano pieces) [17]—are heavily biased toward Western classical and popular styles. The NES Music Database [18] provides 5278 multi-track NES game soundtracks in a custom event format, while POP909 [19] offers 909 piano-based pop arrangements with explicit melody, bridge, and piano tracks for arrangement-generation research. Moving away from keyboard and popular music, the Weimar Jazz Database [20] provides high-quality transcriptions of about 456 monophonic jazz solos with onset-, beat-, and structure-level annotations, and is, therefore, a leading reference point for work on jazz improvisation. At the same time, vocal/folk-oriented symbolic resources such as the Meertens Tune Collections (MTC) [21] document Dutch folksong corpora with scholarly metadata, and community-maintained lead-sheet archives such as Wikifonia (as used in [22]) as well as the Josquin Research Project [23] provide additional symbolic material for melody and early-music research. The MusicNet dataset [24] likewise contributes aligned symbolic annotations for classical-chamber recordings and is often cited alongside MAESTRO in transcription studies.

Dataset construction for non-Western genres has received comparatively little attention. A symbolic Guqin dataset [14] provides 71 pieces in MusicXML, but focuses on a single instrumental tradition rather than the broader traditional-style Chinese melody repertoire. CCMusic [25] aggregates Chinese-music-related datasets for MIR research, including audio resources and unified metadata, but does not provide symbolic note-level representations suitable for melody generation. Crucially, none of the datasets surveyed above is built from Jianpu, which remains the dominant notation format in everyday Chinese music publishing, and none of them addresses the multi-format unification challenge inherent in the present repertoire.

Table 1 summarizes representative symbolic music datasets and positions MSMP within this landscape as a method for heterogeneous melody extraction rather than a static corpus release.

Table 1. Comparison of representative symbolic music datasets and the methodological position of MSMP. "Format" refers to the input modalities accepted; "QC" indicates whether downstream automatic quality control is part of the released pipeline.

Dataset	Size	Genre	Format	QC
Lakh MIDI [8]	176,581 files	Western pop/rock	MIDI	No
MAESTRO [9]	1276 perf.	Classical piano	MIDI + audio	Yes
GiantMIDI-Piano [17]	10,854 pieces	Classical piano	MIDI	Partial
NES Music DB [18]	5278 tracks	Game music	Custom event	No
MusicNet [24]	330 recordings	Classical chamber	MIDI + audio	Partial
POP909 [19]	909 songs	Mandarin pop	MIDI	Partial
Weimar Jazz DB [20]	∼456 solos	Jazz improvis.	SV/MIDI	Manual
Meertens Tune Coll. [21]	∼18 k tunes	Dutch folksong	**kern/MusicXML	Manual
Wikifonia [22]	∼6.5 k sheets	Pop lead sheets	MusicXML	No
Guqin dataset [14]	71 pieces	Guqin music	MusicXML	Manual
CCMusic [25]	Multi-dataset	Chinese (audio)	Audio + meta	Partial
MSMP (ours)	**Method**	**Trad. Chinese**	**Multi-src**	**Auto**

2.2. Optical Music Recognition

OMR research has progressed substantially with deep learning. End-to-end systems such as those surveyed by Calvo-Zaragoza et al. [12] achieve high recognition rates on clean printed Staff notation. The Audiveris engine [16], an open-source Java-based OMR system, provides robust batch processing of Staff images to MusicXML. More recent neural approaches employ sequence-to-sequence models with attention mechanisms to predict music symbols directly from image patches [26].

Nevertheless, Jianpu presents a fundamentally different visual grammar from Staff notation. The pitch alphabet is purely alphanumeric rather than positional: note height on the page carries no tonal information, and rhythmic details depend on underlines (beam-equivalent) and trailing dashes rather than stem and flag geometry. Prior specialized numbered-notation OMR studies [13,27] combine elementary image processing with task-specific learning, and recent expert-system-style approaches with dedicated preprocessing and rule bases have reported substantial progress on printed Jianpu with lyrics [28]. However, such approaches are typically tailored to a single notation family and a single downstream format; to the best of our knowledge, none of them also ingests MIDI and MusicXML within the same pipeline, nor do they expose an explicit melody-oriented quality-control gate of the kind used in Section 3.7. A unified, melody-centred pipeline that bridges across these input types, therefore, remains, in our view, a complementary rather than redundant contribution.

Recent advances in vision language models (VLMs)—including GPT-4V, Gemini, and LLaVA [29]—have demonstrated remarkable generalization to structured document

understanding tasks. By treating the Jianpu score as a spatial-linguistic object rather than an image segmentation problem, VLMs can leverage their pre-trained knowledge of numerical notation and spatial relationships to decode Jianpu without domain-specific training data. This insight motivates the VLM-based transcription module in MSMP.

2.3. Automated Dataset Construction Pipelines

The automation of music dataset curation has been approached through web scraping [30], MIDI synthesis pipelines [8], and score-based digitization projects [31]. The Million Song Dataset [30] pioneered large-scale audio feature extraction but does not provide symbolic representations. Raffel's Lakh MIDI Dataset [8] aggregated MIDI files from online repositories with automatic de-duplication. Vigliensoni, Burlet, and Fujinaga [31] proposed measure-level extraction from scanned score images but stopped short of full music transcription.

A common limitation of existing pipelines is the assumption of a homogeneous input format. The music21 toolkit [15] provides format-agnostic symbolic manipulation and supports MIDI, MusicXML, and Humdrum parsing, but does not address image-based input or automated QC. Our work extends the music21 ecosystem by adding upstream image classification and OMR modules, and downstream QC filtering, to create a complete ingestion pipeline for heterogeneous sources.

To make the positioning of MSMP against *extraction pipelines* (as opposed to *datasets*) more concrete, we also summarize, along five axes, what a small number of directly comparable systems accept as input, what they emit, how they route between notation types, and what (if anything) they do about quality control. Audiveris [16] accepts printed staff images and emits MusicXML with a deterministic rule-based engine; it does not accept Jianpu input, does not perform notation-type routing, and exposes no melody-oriented QC. The end-to-end staff-OMR systems surveyed by Calvo-Zaragoza et al. [12] and the sequence-to-sequence staff transcribers represented by Camera-PrIMuS [26] share the same staff-only input assumption and likewise do not address Jianpu or format-level routing. On the Jianpu side, earlier specialized numbered-notation OMR work [13] and the more recent expert-system approach [28] demonstrate that symbolic-grid Jianpu can be extracted with task-specific pipelines, but these systems operate in isolation from any staff-notation or symbolic-file branch and do not interoperate with MIDI/MusicXML ingestion within the same pipeline. Against this backdrop, MSMP is best characterized as a *systems-integration pipeline* rather than as a new OMR algorithm: it reuses Audiveris on the staff branch and delegates Jianpu transcription to a general-purpose VLM, and its novelty lies in the routing layer, the convergence to a single event-level schema, and an explicit (if *structural*-only) QC gate. We, therefore, do not claim that MSMP transcribes Jianpu more accurately than the specialized approaches of [13,28]; the comparison we can offer within this submission is a within-pipeline alt-prompt ablation on the VLM Jianpu branch (Section 4.7), and a direct head-to-head against a specialized Jianpu OMR system is explicitly listed as future work in Section 6.

3. Materials and Methods

3.1. Data Sources

The prototype material used in this study was assembled from two publicly accessible Chinese-language online score repositories that provide a large volume of free Jianpu pages: *qupu123.com* (https://www.qupu123.com/, accessed on 20 April 2026) and *2qupu.com* (https://www.2qupu.com/, accessed on 20 April 2026). Within these portals, the pipeline specifically targeted the Jianpu-oriented sections `/jipu` and, where available, `/yuanchuang` and individual-contributor spaces, following the link structure exposed on

each site. The resulting pages cover folk songs, revolutionary-era songs, film/TV music, and vocal works that are typical for present-day Chinese music consumption. Images were downloaded in JPEG and PNG format at their native resolutions (ranging from 400×600 to 2400×3200 pixels). For methodological validation, the collection was stored locally and organized by source website with standardized file naming; the pipeline records the exact URL of each retrieved page so that the origin of every extracted segment can be traced back. Rendered staff-notation pages from the same two portals were also included when available, so that the Staff-OMR branch could be exercised on real-world layouts.

In principle, MSMP also accepts MIDI and MusicXML files; however, the present case study focuses on image-based sources because Jianpu remains the most practically important extraction target in current circulation, and because, to the best of our knowledge, no well-known public MIDI or MusicXML corpus of traditional-style Chinese melodies exists at a scale comparable to the image material available on these portals.

3.2. Incremental Collection Protocol

The image-collection component was designed as an incremental rather than one-shot harvesting procedure. In addition to local score files already present in the initial repository, the orchestration script can issue a rotating list of source queries to download additional Jianpu pages from public score websites. The default query pool combines source-aware terms with broader music-oriented queries so that the automated collection script can continue harvesting when a single source becomes saturated.

Two early-stopping criteria are used to keep the process operationally efficient. First, the download stage stops after consecutive rounds with no newly downloaded images. Second, the processing stage stops after consecutive rounds in which the size of the standardized melody corpus no longer increases. This design reflects the practical goal of incremental corpus growth under unstable web availability rather than exhaustive web crawling.

To support reproducibility and post hoc analysis, the pipeline records collection progress, per-query download statistics, and item-level processing outcomes in structured diagnostic logs. These records make it possible to analyze not only the final corpus, but also the dynamics of data acquisition and the yield of each processing round.

3.3. System Architecture

Figure 1 summarizes the overall architecture of MSMP, which combines parallel ingestion with standardized convergence so that heterogeneous inputs can be processed by notation-appropriate modules before entering a common validation layer.

The architecture begins with two parallel input branches. Born-digital symbolic files are parsed and normalized with `music21`, whereas scanned score images are first routed by a zero-shot Vision-Language Model (VLM) and then dispatched to notation-specific backends: Audiveris for staff notation and Gemini-2.5 for Jianpu transcription. The outputs of both branches are subsequently mapped to a unified step-indexed event schema, allowing downstream processing to remain independent of source format. In the final stage, candidate segments are filtered by a melody-oriented quality control module and retained only when they satisfy structural validity requirements; diagnostic statistics are recorded in parallel to support later error analysis and corpus-level evaluation.

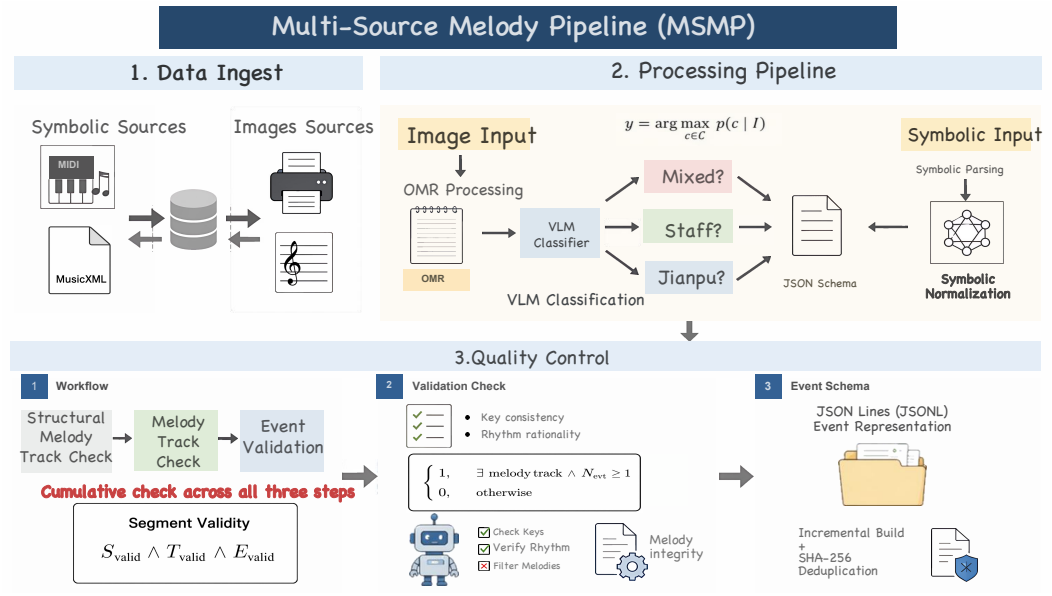


Figure 1. Schematic overview of the MSMP architecture. The framework consists of three stages: (1) **Data Ingestion**, where heterogeneous sources (MIDI, MusicXML, and raw scanned images) are collected; (2) **Processing Pipeline**, which splits into a symbolic parsing branch for digital files and an OMR branch for score images. The OMR branch employs a Vision-Language Model (VLM) for zero-shot classification, routing Jianpu to a VLM-based transcriber and standard staff to a rule-based engine, before converging into a unified JSON schema; (3) **Quality Control**, which performs structural and melody-track validation before serializing the final standardized Event Schema.

3.4. Score-Type Classification

The score-type classifier is implemented as a single-turn VLM inference call. The system prompt provides brief descriptions of the four target categories:

> *"Classify the following sheet music image into exactly one category: 'jianpu' (numbered notation using digits 1–7), 'staff' (standard five-line notation), 'mixed' (both systems on the same page), or an invalid category such as lyrics-only or unreadable noise. Return a JSON object with a single field 'type'."*

Because the prototype corpus was harvested exclusively from the Jianpu-oriented sections of two Chinese-language portals (*qupu123.com/jipu* and *2qupu.com*; see Section 3.1), the classifier in this study is exercised almost exclusively on pages that are *a priori* Jianpu by source. We, therefore, treat the classifier as a lightweight sanity-check/defence-in-depth layer rather than as a balanced four-class decision module: it is responsible for diverting the rare off-category image that slips through source-directory filtering, not for deciding between populations of equal size. We have *not* benchmarked the classifier on a curated, labelled staff/mixed/invalid test set in the present submission, because the corpus itself does not contain balanced staff or mixed samples; producing such a labelled out-of-distribution test set is explicitly listed in Section 6 as future work. A small, honest indication of routing behaviour can, however, be extracted from the manually annotated benchmark reported later in Section 4.7: 48 of the 50 annotated pages are printed Jianpu and were routed to the Jianpu VLM branch (correct routing); two of the 50 pages, despite being retrieved from the Jianpu subsection, turned out on manual inspection to be printed in standard staff notation, and the classifier did *not* catch these two cases—it forwarded them to the Jianpu branch, where they contributed to the note-level errors reported in Section 4.7. We report this as case-study evidence of a routing-stage failure mode rather than as a generalisable accuracy number, and we acknowledge in Section 5 that producing

a proper confusion matrix is the correct next step. The full classifier and transcription prompt templates are reproduced in Appendix A for reproducibility.

3.5. Optical Music Recognition Strategy

3.5.1. Staff Notation

Images classified as *staff* are processed using the open-source Audiveris engine [16], invoked via its batch export interface. This module generates standard MusicXML files, which are subsequently validated for XML well-formedness and ingested into the pipeline's symbolic parsing stage. While Audiveris provides a robust baseline for standard western notation, it inherently faces challenges with complex score layouts, such as dense orchestral pages or irregular beam groupings. In cases where the engine fails to produce valid XML output or crashes due to image noise, the system logs the failure and excludes the page from the final dataset, prioritizing high-precision extraction over partial recovery of corrupted scores.

3.5.2. Jianpu Notation

Jianpu, or numbered musical notation, serves as the dominant medium for Chinese popular music distribution but lacks a widely adopted OMR standard. Its visual grammar differs fundamentally from staff notation, relying on a relative pitch system where the seven diatonic scale degrees are represented by Arabic numerals (1–7). Octave shifts are denoted by dots placed above (higher) or below (lower) the digits, while rhythmic duration is encoded through graphical affixes: unmodified numerals represent quarter notes, underlines indicate halving of duration (e.g., one underline for an eighth note), and trailing dashes extend it. Key signatures are explicitly declared as textual headers (e.g., $1 = D$), and additional articulation marks such as slurs and dynamics are superimposed on this numeric grid.

To address this unique syntax, we leverage the spatial-reasoning capabilities of *Gemini-2.5-flash* [32]. The choice of this particular VLM, rather than an open-source or locally hosted model such as Qwen-VL or InternVL, was driven by three practical considerations during the prototype phase: (i) Gemini-2.5-flash offered, at the time of the experiments, a consistently strong zero-shot behaviour on mixed text–image document understanding tasks, which is the closest analogue to Jianpu transcription; (ii) it provides a long-context JSON-friendly output mode through the same OpenAI-compatible endpoint used by the rest of the pipeline, avoiding an additional inference stack; and (iii) for a methodological prototype focused on *whether* a general-purpose VLM can take the role of a dedicated Jianpu OMR engine, using a widely accessible hosted model removes most of the engineering overhead of fine-tuning, GPU provisioning, and model selection. We note explicitly that this choice also introduces reproducibility and privacy limitations that motivate the planned migration to a fine-tuned open-source local model (see Section 6); the pipeline architecture itself is deliberately model-agnostic.

Instead of a direct image-to-MIDI translation, which often hallucinates alignment, the system uses a structured prompt that enforces a notation-aware intermediate representation. The model is instructed to first extract global metadata—including the key signature and time signature—and then traverse the score linearly to transcribe note events. The output is constrained to a rigorous JSON schema where notes are encoded with explicit Jianpu tokens (e.g., `1^` for a high C, `5_` for a low G eighth note) and precise beat durations. This intermediate serialization allows the subsequent normalization module to deterministically map relative pitches to absolute MIDI numbers and align rhythmic sequences without ambiguity, effectively bridging the gap between visual notation and computational representation.

3.5.3. Error Handling and Retry Logic

The pipeline enforces reliability through a lightweight but strict validation protocol. Post-inference checks ensure that the VLM output conforms to the expected JSON structure and contains valid musical hierarchies. Pages that fail these structural checks—or those that produce mathematically impossible rhythm sequences—are discarded immediately rather than attempting heuristic repair. To optimize computational resources, successful OMR results are cached by image hash, ensuring that the computationally expensive VLM inference is performed only once per unique image across multiple experimental runs.

3.6. Unified Event Schema

To ensure interoperability across heterogeneous sources, all ingestion branches—whether parsing symbolic files or transcribing images—converge on a single, standardized event schema. This schema is designed to be minimal yet sufficient for melodic analysis, discarding notation-specific formatting while preserving essential musical information.

The core data structure is organized hierarchically. At the top level, each item is identified by versioning and segment identifiers, accompanied by global metadata that includes the key signature, time signature, and source provenance. Temporal information is normalized into a grid-based system, defining parameters such as steps per beat, beats per bar, and global tempo. This grid ensures that rhythmic values from different notation systems (e.g., Jianpu's relative duration markers versus MusicXML's division-based timing) are mapped to a common, mathematically consistent timeline.

Musical content is represented as a sequence of discrete events. Each segment contains track descriptors with semantic role labels where available, together with an ordered sequence of note-on and note-off operations defined by step index, track identity, MIDI pitch, and velocity. The record further stores validation metadata and summary metrics computed during quality control, such as melody validity checks and polyphony rates. By adopting this low-level, step-indexed representation, MSMP decouples the complex upstream task of notation parsing from downstream applications, facilitating uniform analysis and generation tasks regardless of data origin.

3.7. Structural Quality Control

The acceptance layer of the present implementation is, by design, a *structural* Quality Control (QC) gate rather than a content-level accuracy filter. A candidate segment is accepted if two conditions are met: (1) the normalized track list contains at least one track identified as melodic; and (2) the segment contains at least one musical event. Accepted segments are, therefore, assigned structural validity metadata and retained as melody-centered material in the final corpus. We stress this wording explicitly because the term "quality control" can be misread as implying musical-correctness filtering; in our framework, the gate is only responsible for rejecting segments whose structure cannot be consumed by downstream tooling, and an accepted segment is not, by that fact alone, guaranteed to be a faithful transcription of the source page.

In addition to this acceptance decision, the schema retains auxiliary summary fields such as melody polyphony rate, out-of-scale rate, accompaniment coverage, and pair-validity indicators. In the present prototype, these fields function primarily as extensibility hooks and diagnostic descriptors; the active acceptance policy remains intentionally lightweight. Content-level validation—for example, note-by-note comparison against a manually annotated reference—is handled as a separate, offline measurement (Section 4.7) rather than as a filter, because any content-aware filter that depends on the model's own output would be circular. This design keeps the extraction method simple and robust while

preserving room for stricter downstream filtering—including the residual-error-aware triage sketched in Section 5—in future versions.

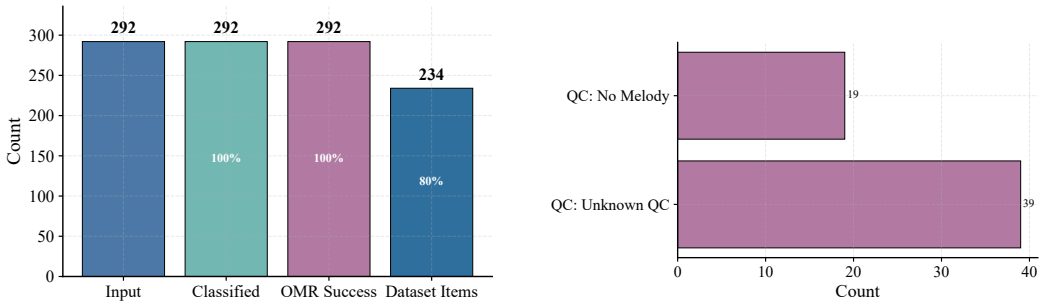
3.8. Implementation Details

The implementation combines established symbolic parsing and OMR components with a VLM-based transcription module. Symbolic files are processed with `music21`, staff notation is handled by Audiveris, and Jianpu pages are transcribed with Gemini-2.5-flash. To improve stability and reproducibility, the orchestration layer uses conservative inference settings, moderate parallel execution, and content-based deduplication of previously processed inputs. Runtime, therefore, reflects the combined effects of data acquisition, cache reuse, model latency, and transcription failure rather than an isolated model-only benchmark.

4. Results

4.1. Prototype Workflow Validation

The prototype study confirms that MSMP can execute the full extraction loop from raw score images to standardized symbolic output. In the prototype run, score images were first normalized by directory structure and filename conventions before entering the pipeline; the score-type classifier then provided a practically adequate separation between printed Jianpu pages and other notation types, enabling automatic backend selection. Pages routed to the VLM branch yielded JSON outputs that encode key, meter, and note-event sequences in a form suitable for downstream normalization. Finally, segments that passed the melody-oriented QC gate were serialized into a uniform event-level representation with explicit temporal, track-role, and validation metadata. A more quantitative breakdown of the per-stage attrition is reported in Figure 2.



(**a**) Pipeline funnel. (**b**) Failure breakdown.

Figure 2. (**a**) Pipeline funnel retention rates for the validation cohort; (**b**) failure breakdown by stage and reason.

4.2. Collection Dynamics and Processing Yield

To complement the musicological statistics, we also inspected the operational logs generated by the pipeline during collection and batch processing. A representative prototype run exhibited a strongly front-loaded acquisition pattern: most newly usable material was obtained in the first query round, whereas subsequent rounds produced no additional images and, therefore, triggered early stopping. Concretely, the local image pool grew from 271 to 316 files during the first source-specific Jianpu-related query round (an increase of 45 files after 28 recorded successful downloads), while two subsequent rounds with broader notation-related queries produced no further image growth. The first processing round then lifted the non-augmented raw-record count from 256 to 610 (a gain of 354 records), after which two additional rounds yielded no new raw records and the orchestration

logic triggered an early-stopping rule because both image growth and raw-yield growth had stagnated.

Reconciliation of image, record, page, and segment counts. Because the numbers quoted in this subsection and in Section 4.3 come from different stages of the same pipeline, we make their relationship explicit here so that readers can verify the funnel end-to-end. The *image pool* is the set of distinct source image files on disk (316 files after early stopping). The *raw-record* count (610 after round 1) counts every VLM transcription record that was serialised to `raw.jsonl` across all processed images: because a single multi-page score or multi-song page can produce more than one transcription record, raw-record count is larger than image-pool size and is not a per-page count. The *validation cohort* of 292 *pages* is the slice of the raw-record stream that was re-processed under the fixed configuration used for the quantitative results in Sections 4.3, 4.6 and 4.7; one page corresponds to one image file that was successfully pre-processed and that reached the OMR stage. Of these 292 pages, 234 (80.1%) passed both the OMR stage and the structural QC gate and were retained in the *accepted-page* set, while 58 pages were rejected for the reasons broken down in Figure 2b. An accepted page is then split by the structural QC gate into one or more *accepted melody segments* (a segment is the unit that downstream musicological statistics and the ground-truth benchmark are indexed against), so segment count is, by construction, no smaller than accepted-page count. The 50-page manually annotated ground-truth subset of Section 4.7 and the 10-page self-consistency subset of Table 4 are drawn from this accepted-page set by stratified sampling on notation family and, therefore, they do not add new raw data to the funnel—they only re-label a portion of it for quantitative evaluation. We emphasise that this funnel describes one prototype run: as noted in Section 5, the per-round growth figures are *system-behaviour observations* of that run, not stable properties of the repertoire source space.

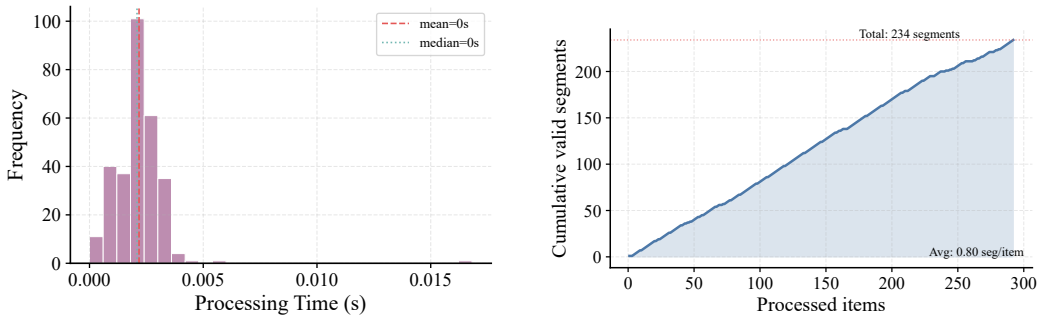
4.3. Pipeline Retention, Failure, and Runtime

Figure 2a illustrates the processing yield on a validation cohort of 292 pages. In this run, 234 pages (80.1%) yielded at least one accepted melody segment after passing both the OMR and QC stages. This retention rate reflects the combined efficacy of the VLM transcriber and the melody-oriented filter.

The failures are not concentrated in a single mechanism. As shown in Figure 2b, some pages failed because the VLM branch returned malformed JSON or timed out, whereas others completed transcription but did not yield an output compatible with the current melody-oriented acceptance gate. This distinction is important because it separates front-end recognition failures from downstream normalization or role-label failures.

Runtime also exhibited a heavy-tailed, bimodal distribution (Figure 3a). The values plotted are per-item wall-clock seconds as recorded in `metrics.jsonl`, i.e., the wall time from picking up a source file to writing its standardized JSON; they, therefore, include data acquisition, cache look-up, VLM call, JSON parsing, normalization, and QC, but exclude the time spent rendering the final figures and PDF report. Two regimes are clearly visible on the plot. A first, dense cluster sits below 1 s: these are items whose OMR output had already been produced in an earlier run and was retrieved from the image-hash cache, so no new VLM call was issued. The remaining items correspond to first-time VLM calls on previously unseen pages; in this subpopulation the mean processing time was 346.1 s and the 90th percentile exceeded 1222 s, dominated by long-latency VLM calls, retries, and a small number of particularly difficult pages. The overall mean and 90th percentile reported on the figure (346.1 s and 1222 s) are, therefore, *aggregate* statistics across both regimes, not the typical time for a single VLM inference when the cache is hit. This result reinforces that

end-to-end processing time should be interpreted as an operational property of the whole pipeline under a given caching state, not merely as model inference time.



(**a**) Per-item processing time. (**b**) Cumulative valid-segment yield.

Figure 3. (**a**) Per-item processing time in the representative run (seconds, linear *x*-axis). The histogram is *bimodal*: the very tall bar at the left edge (<1 s) is the cache-hit sub-population (no VLM call was issued) and the long right tail above a few hundred seconds is the first-time VLM sub-population. The mean (346.1 s) and 90th percentile (1222 s) annotations on the plot are *aggregate* statistics over both regimes; they do *not* describe a typical cache-hit item, which is sub-second, nor the centre of the VLM-call sub-population alone, which sits around 400 s. (**b**) Cumulative valid-segment yield across processed items.

The cumulative-yield curve in Figure 3b is close to linear, with a mild sub-linear drift in the second half of the run. In other words, in this validation cohort the pipeline keeps producing new accepted segments at a roughly constant rate rather than saturating sharply: once a page passes QC, it typically contributes a small number of segments regardless of its position in the queue. The small gap between the observed curve and a purely linear reference is attributable to the 19.9% of pages that failed QC (see Figure 2), which are distributed fairly uniformly over processing order and, therefore, produce short horizontal plateaus rather than a pronounced elbow.

4.4. QC Ablation Study

To examine the practical role of the melody-oriented QC layer, we performed a controlled ablation on a cached subset of 292 pages, comparing *QC on* and *QC off* under identical upstream OMR conditions. Disabling QC increased the number of accepted pages from 234 to 253, a relative gain of 7.5%. This gain comes from retaining outputs that the current melody gate would otherwise reject. However, the medians of the four inspected quality descriptors—note density, leap rate, rhythm variability, and pitch entropy—remained similar between the two conditions: note density changed from 1.388 to 1.376 notes per beat, rhythm coefficient of variation (CV) from 0.595 to 0.605, and pitch entropy from 2.409 to 2.429 bits. We, therefore, interpret the current QC layer as a conservative acceptance mechanism that removes a limited number of structurally ambiguous pages rather than as a strong style-shaping filter.

4.5. Qualitative Case Studies

Figure 4 presents three representative pages: an accepted example, a schema-mismatch example, and a failure example. The accepted page produced a structured representation that could be normalized directly into the event schema. The second page illustrates a more subtle error mode: the VLM output was musically plausible, but it followed an internal organization not supported by the current normalization stage. The third page demonstrates a harder failure mode in which a dense layout led to severe transcription breakdown.

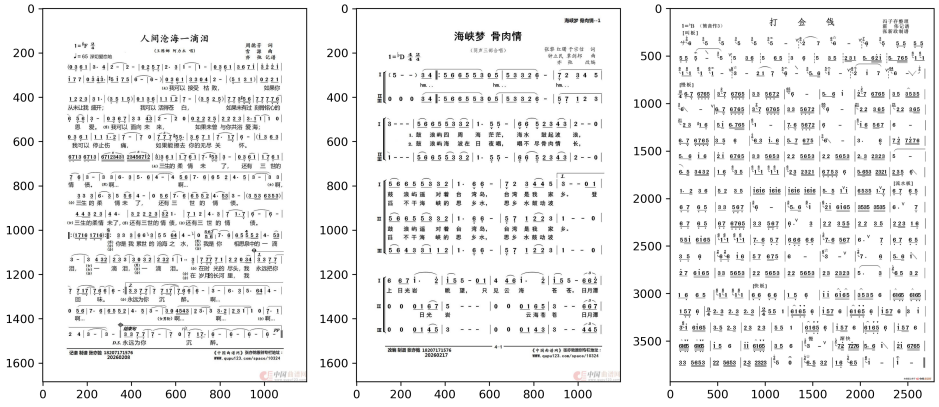


Figure 4. Representative qualitative cases from the Jianpu branch. (**Left**) accepted page. (**Middle**) musically plausible but schema-mismatched output. (**Right**) failure case associated with severe transcription breakdown.

These qualitative cases are useful because they show that current failure modes are not limited to obviously unreadable scans. Some pages fail because the model produces a structurally different but still musically meaningful JSON organization. This observation directly motivates stronger post-processing, schema adapters, and finer-grained error taxonomies in future versions of the pipeline.

4.6. Musicological Validation

The engineering evaluations above do not establish note-level accuracy, but they do show that the extracted outputs are sufficiently coherent to support downstream symbolic analysis. Figure 5 presents the joint distribution of musical keys and time signatures over the 234 accepted segments in the validation cohort. For time signatures, 4/4 accounts for 182 segments (77.8%), followed by 2/4 with 36 segments (15.4%), 6/8 with 14 segments (6.0%), and 3/4 with 2 segments (0.9%). The distribution of declared keys (Figure 5a) is more spread out, with 213 segments providing an explicit key label and a long tail of sharp and flat keys; the three most frequent bins are F (44 segments, 20.7%), D (28 segments, 13.1%), and C (23 segments, 10.8%). All percentages and counts in this paragraph are read directly from the underlying statistics file used to produce Figure 5, so the numbers in the figure and in the text now match by construction. This tonal concentration—a clear preference for sharp/flat-free or single-flat keys together with a handful of $1 = \mathrm{D}$ and $1 = \mathrm{E}$ arrangements—is consistent with the fact that $1 = \mathrm{C}$, $1 = \mathrm{D}$, $1 = \mathrm{F}$, and $1 = \mathrm{G}$ are the most commonly declared Jianpu key signatures in amateur and vocal repertoires.

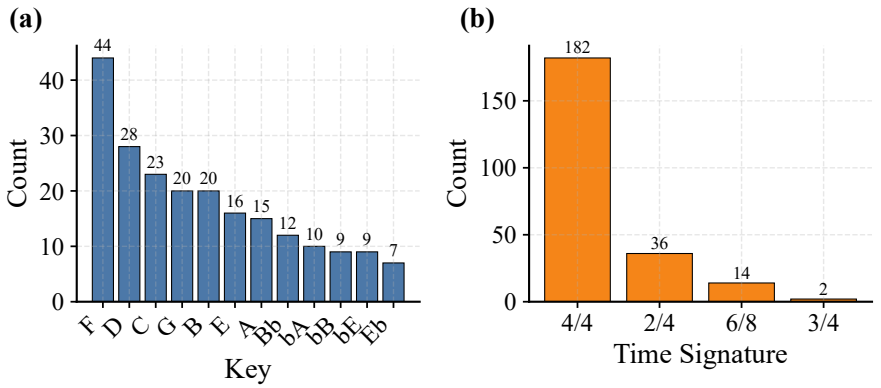


Figure 5. Distribution of musical keys (**a**) and time signatures (**b**) in the prototype melody corpus extracted by MSMP. Bar heights are absolute counts; percentages in the text are computed from the same counts.

Figure 6 shows that pitch range is centered on a median of 20.5 semitones, while note durations cluster strongly around eighth-note (0.5 beats) and quarter-note (1 beat) values. The resulting tessitura and rhythmic granularity are consistent with vocal and single-line instrumental writing.

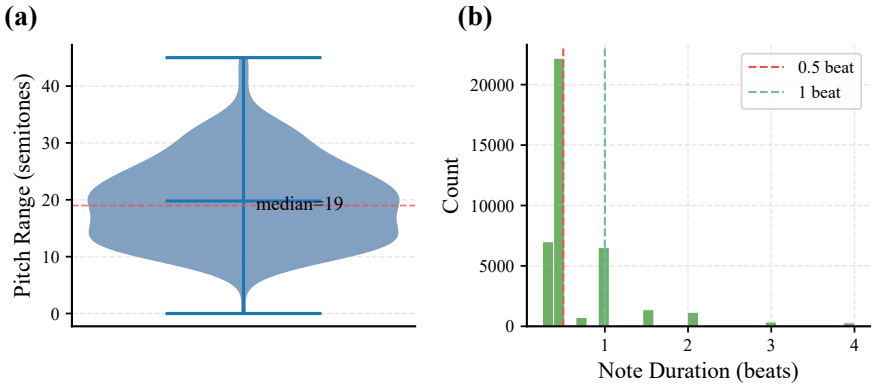


Figure 6. Pitch range per segment (**a**) and note-duration distribution (**b**) in the prototype melody corpus.

The interval transition matrix in Figure 7 provides one additional, purely distributional view of the extracted melodies. The matrix shows a clear concentration of probability mass in the neighbourhood of small signed intervals, consistent with the general tendency of tonal melodies to favour stepwise motion [33]; at the same time, off-diagonal structure is also clearly present, reflecting leaps and direction changes that are typical of the source material. We, therefore, read Figure 7 as evidence that short-interval continuation is *more frequent than*, but does not fully dominate, larger-interval transitions, and we do not claim that the extracted melodies are "smooth" in a stronger sense.

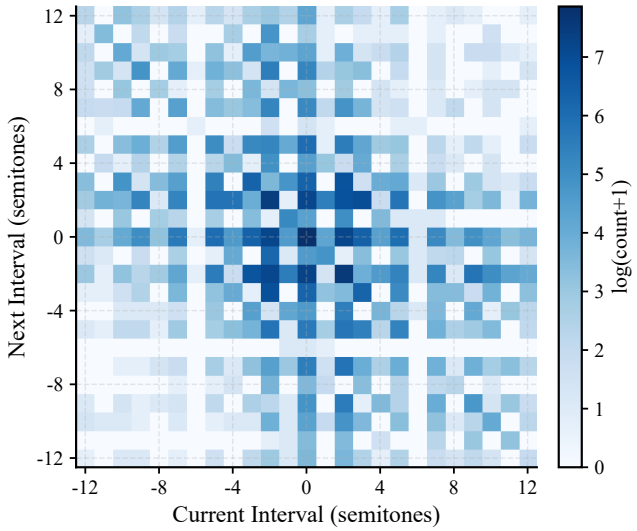
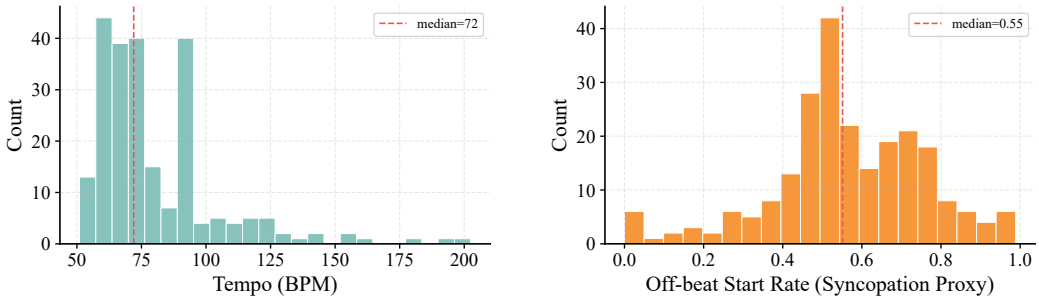


Figure 7. Interval transition matrix computed over the prototype melody corpus. Colour encodes the empirical transition probability $P(\text{interval}_{t+1} \mid \text{interval}_t)$; rows and columns are semitone signed intervals centred on 0.

We additionally report two robust distributional descriptors of rhythmic behaviour that are used in the QC ablation (Section 3.7): the large-leap rate and the rhythm coefficient of variation (rhythm CV). Across the accepted segments, the median large-leap rate is about 0.005 and the median rhythm CV is about 0.66, which is in line with melody-centred single-line writing and with the ablation-level medians reported above (0.595–0.605 depending on

the QC setting). These numbers are presented here purely to document the behaviour of the extraction pipeline on real input; they are not intended as musical-quality scores.

Additional descriptors remain consistent with melody-centered symbolic material. Tempo is right-skewed with a median of 70 Beats Per Minute (BPM) (Figure 8a), and the median off-beat start rate is 0.52 (Figure 8b), indicating that the pipeline preserves sub-beat rhythmic structure rather than collapsing events onto downbeats.



(**a**) Tempo (BPM) distribution. (**b**) Off-beat note start rates.

Figure 8. Additional descriptors: (**a**) tempo distribution; (**b**) off-beat note start rates.

4.7. Ground-Truth Benchmark on a Manually Annotated Jianpu Subset

The distributional evidence above characterizes the output of the pipeline but does not verify its fidelity to individual source pages. To address this concern directly, and in response to the second reviewer's observation that structural validity alone cannot separate faithful transcriptions from "plausible-looking" ones, we constructed a manually annotated Jianpu benchmark of 50 pages and ran a quantitative, per-page comparison against the pipeline's cached outputs.

Benchmark construction. The 50 pages were selected by evenly-spaced deterministic sampling (fixed random seed) from the 292-page cached cohort used throughout Section 4.3, so that benchmark pages and ablation pages come from the same representative population. For each page we manually read the printed header and first musical line and recorded: (i) the declared key signature (e.g., $1 = {}^{\#}\mathrm{F}$, $1 = \mathrm{G}$, $1 = {}^{\flat}\mathrm{E}$); (ii) the declared time signature (e.g., 4/4, 2/4, 6/8, 3/4); (iii) the declared tempo in beats per minute whenever it was printed; and, for ten pages, (iv) the first sixteen melodic note events (pitch symbol and duration). Three pages had no printed key/meter/tempo markings and were excluded from the corresponding headline metrics; two additional pages were printed in standard staff notation but had nevertheless been routed to the Jianpu branch by the upstream classifier, and we retained them in the benchmark precisely because they stress-test the notation-aware router.

Metrics. On this benchmark we report five complementary metrics. *Key (tonal-pitch-class) accuracy* treats $1 = {}^{\flat}\mathrm{B}$ and $1 = \mathrm{Bb}$ as the same key, which is the musicologically correct comparison across spelling conventions. *Key (literal spelling) accuracy* in addition requires exact string match, which exposes spelling-convention drift. *Time-signature exact accuracy* requires the string p/q to match. *BPM* $\pm 1/\pm 5$ report the fraction of pages whose predicted tempo is within one or five beats per minute of the printed value. For the ten pages with note-level annotations we additionally compute a pitch-code F1 (treating the first sixteen notes as a multiset of pitch codes, counting octave dots), a pitch-class F1 (octave-invariant), and a Symbol Error Rate based on a Levenshtein edit distance between the predicted and reference pitch-code sequences.

Results. Table 2 summarizes the outcome. Time signature and tempo are recovered with high reliability: time-signature exact accuracy is 95.8% and every sample with a printed BPM is predicted within ± 1 of the reference. Key is recovered at 77.1% when enharmonic

spellings are treated as equivalent, and at 66.7% when exact spelling is enforced. The dominant error pattern is an *accidental-dropping* failure in which the pipeline returns, e.g., 1 =F where the printed marking is 1 = ♯F, or 1 =D in place of 1 =♭ D; this failure mode is exactly the "plausible-looking output" case flagged by reviewer 2, because the surrounding JSON structure and the note sequence both parse cleanly. On the 10 note-level subset, first-16-note pitch F1 is 0.898 at exact pitch-code granularity and 0.955 under octave-invariant comparison, and the mean Symbol Error Rate on the pitch sequence is 0.150. We interpret these numbers as a first *honest* accuracy estimate of the current Jianpu branch, not as a claim of production-grade correctness. The gap between 95.8% (time signature) and 66.7% (literal key spelling) in particular indicates that the weakest link is key/accidental reading, which will receive targeted attention in future revisions (Section 5).

Table 2. Ground-truth benchmark on 50 manually annotated Jianpu pages from the prototype cohort. n is the number of pages contributing to each metric after exclusions.

Metric	Definition	Value	n
Key (pitch class)	Enharmonic-equivalent tonal pitch class match	77.1%	48
Key (literal spelling)	Exact string match (e.g., ♭B vs. B♭ distinct)	66.7%	48
Time signature	Exact p/q match	95.8%	48
BPM within ± 1	$\lvert\hat{b} - b\rvert \leq 1$	100.0%	44
BPM within ± 5	$\lvert\hat{b} - b\rvert \leq 5$	100.0%	44
Pitch F1 (first 16 notes)	Multiset F1 on integer pitch codes	0.898	10
Pitch-class F1 (first 16)	Multiset F1, octave-invariant	0.955	10
Symbol Error Rate	Mean Levenshtein/$\lvert$ref$\rvert$	0.150	10

Error structure. A breakdown of key errors on this benchmark shows that 8 of 11 literal-spelling mismatches are accidental-dropping (flat or sharp omitted from a correct letter), two occur on pages printed in standard staff notation that were mis-routed to the Jianpu branch at the classifier stage, and the remaining one is a mode-level confusion on a SATB staff page that was similarly mis-routed. Time-signature errors are 2 of 48 and both collapse a 2/4 page to 4/4, consistent with downbeat-grid smoothing rather than arbitrary hallucination.

Token-level side-by-side illustrations. To make the failure modes above concrete rather than merely statistical, Table 3 shows, for two representative Jianpu songs from the 10-page note-level subset, the printed reference and the pipeline prediction placed side-by-side for the first 16 melodic events. The first song (Music 1) is the canonical *accidental-dropping* case: the printed key is 1 =♯F and the printed meter is 2/4, whereas the pipeline returns 1 = F and 4/4; the printed first-16-note sequence ends on $7-\hat{1}-7-7$ (octave-dot above the 1), but the pipeline drops the octave dot on the 14th token and then further inserts a $2-3$ tail instead of reproducing the held $7-7$ cadence, which is exactly the kind of structurally valid yet content-inaccurate output that motivates a content-level QC layer (Section 5). The second song (Music 2) is a positive control on the same 1 =♯F key: the pipeline also drops the sharp in the key signature, but the first-16-note pitch sequence is recovered exactly at pitch-class granularity, so the page contributes a 0/16 token-level mismatch to the pitch F1 figure of Table 2 while still being counted as a key-spelling error in the same table. These two rows together illustrate how the 77.1% pitch-class/66.7% literal split for key and the 0.898/0.955 pitch F1 split emerge in practice: metadata errors (dropped accidental, 2/4→4/4) and note-level errors are largely decoupled per page, and the dominant note-level errors are local—octave-dot or splitting decisions on a small number of tokens—rather than global hallucinations of the melody.

Table 3. Token-level side-by-side comparison of ground truth vs. pipeline prediction for the first 16 melodic events on two representative pages from the 10-page note-level subset. Cells with a differing prediction are underlined; the octave-up mark is written as $\hat{1}$ to match the Jianpu convention of Appendix A.

Page	Field	Value
Music 1	Key (GT/Pred)	$1 = {}^{\#}\mathrm{F}$/$\underline{1 = \mathrm{F}}$ (accidental dropped)
	Meter (GT/Pred)	2/4/$\underline{4/4}$
	GT first-16	0, 3, 6, 1, 3, 4, 2, 2, 0, 0, 2, 5, 7, $\hat{1}$, 7, 7
	Pred first-16	0, 3, 6, 1, 3, 4, 2, 2, 0, 0, 2, 5, 7,$\underline{2}$, $\underline{2}$,$\underline{3}$
Music 2	Key (GT/Pred)	$1 = {}^{\#}\mathrm{F}$/$\underline{1 = \mathrm{F}}$ (accidental dropped)
	Meter (GT/Pred)	2/4/2/4
	GT first-16	1, 1, 2, 2, 1, 3, 2, 3, 0, 2, 2, 6, 6, 5, 5, 0
	Pred first-16	1, 1, 2, 2, 1, 3, 2, 3, 0, 2, 2, 6, 6, 5, 5, 0 (pitch-class exact)

Self-consistency audit. Agreement against ground truth is one axis of reliability; stability under repetition is another. Because the Jianpu branch is driven by a non-zero-temperature VLM, two repeated calls on the same image need not produce bit-identical outputs, and Reviewer 2's concern about "plausible-looking" outputs could in principle reflect either systematic content errors or merely per-call noise. To separate the two effects we performed a small self-consistency audit. Ten pages drawn from the same cohort were each queried $K = 3$ times with the identical prompt, yielding 30 VLM calls in total; 26 of them (86.7%) returned a structurally valid JSON, the remaining four failed at the network layer after the configured retry budget was exhausted. On the 22 pairs of runs that can be compared (all ten images retained at least two successful runs), we computed pair-wise agreement on the declared key, time signature, and tempo, as well as pair-wise Pitch F1 on the first-16-note pitch codes using the same metric definitions as the ground-truth comparison above. Results are summarised in Table 4.

Table 4. Self-consistency of the Jianpu VLM branch on 10 pages, $K = 3$ runs per page, 22 successful run-pairs in total.

Quantity	Definition	Value
Successful calls	Out of $K \times$ pages $= 30$ total calls	26/30 (86.7%)
Key agree rate	Fraction of run-pairs with identical predicted key	22/22 (100.0%)
Time-signature agree rate	Fraction of run-pairs with identical p/q	22/22 (100.0%)
BPM agree rate (exact)	Fraction of run-pairs with identical predicted BPM	22/22 (100.0%)
First-16 exact match	Fraction of run-pairs with bit-identical pitch sequence	0.136
First-16 pitch-code F_1	Multiset F_1 over the first-16 pitch codes	0.799
First-16 pitch-class F_1	Octave-invariant multiset F_1	0.930

Two observations follow. First, the metadata fields (key, time signature, BPM) are *deterministic across runs*: every one of the 22 run-pairs agrees exactly on all three. This implies that the accidental-dropping and 2/4→4/4 errors reported above are not artefacts of per-call randomness but systematic reading errors that the VLM makes in the same way every time; they are, therefore, well-defined targets for a future content-level QC layer (Section 5), rather than noise that would average out under majority voting. Second, the first-16-note pitch sequence is only bit-identical on 13.6% of the run-pairs, but multiset Pitch

F1 remains 0.799 at exact pitch-code granularity and 0.930 once octave dots are folded out. Inspection of the mismatching pairs shows that they differ mostly in octave-dot placement and in the exact splitting of a dotted pattern (e.g., whether an eighth is written as `0.5` or as `0.25 + 0.25`), not in the underlying pitch alphabet, which is consistent with the 0.150 mean SER reported above and with the Pitch F1 figures in Table 2. The self-consistency audit, therefore, does not replace the ground-truth benchmark, but it provides a useful lower bound on how much of the note-level error can in principle be attributed to per-call VLM variance rather than to content-level mis-reading.

4.8. Alternative-Prompting Baseline on the Note-Level Subset

A legitimate concern about a single-prompt VLM pipeline is that the numbers reported above could be an artefact of the particular prompt in Appendix A rather than a property of the underlying VLM capability. Reviewer 3 specifically asked for a comparison against "at least a simple alternative prompting strategy." Because a side-by-side run against a specialised Jianpu-OMR system was out of scope for this revision (see Section 5 and the *claims and non-claims* list in Section 1, we instead ran a controlled *alternative-prompting* study on the same 10-page note-level subset used in Table 2. We kept the VLM (Gemini-2.5-flash), the images, the ground-truth annotations, and every metric definition unchanged, and varied only the prompt.

Three prompts were compared, results are summarized in Table 5:

- **Original**: the full schema-anchored prompt of Appendix A, which fixes the JSON field names (`key`, `time_signature`, `bpm`, `parts`→`measures`→`note/duration`) and the Jianpu conventions (octave dots written as `1^`, `6_`; lengths anchored to a quarter note).
- **Minimal**: a one-sentence Chinese instruction that says "this is a Jianpu image; return a JSON with fields `key`, `time_signature`, `bpm`, `parts` (each with `role` and `measures`; each note has `note` and `duration`)". No Jianpu conventions, no octave-dot syntax, no example.
- **Chain-of-thought (CoT)**: a two-step prompt that first asks the VLM to describe the page in natural language (key, meter, tempo, then note-by-note the first line of the melody), and only then asks it to emit the JSON.

All three prompts were run at the same model temperature and timeout settings as the main pipeline.

Table 5. Alternative-prompting baseline on the same 10-page note-level subset as Table 2. Model and images are identical across rows; only the prompt varies. n_{ok} is the number of pages that returned a structurally valid JSON (after at most five network-level retries per page). Pitch F1 and SER use the same definitions as in Section 4.7.

Prompt	Design	n_{ok}	Key (pc)	Pitch F1 (pc)	SER
Original (Appendix A)	Full schema + Jianpu convention + example	10	1.000	0.931	0.156
Minimal	One-sentence field list; no Jianpu convention	10	1.000	0.945	0.135
Chain-of-thought	Describe-then-JSON, two-step	10	0.900	0.706	0.413

Three observations follow.

First, on the *metadata* axis the prompt matters very little: both **Original** and **Minimal** recover the key at pitch-class granularity on every structurally successful page, and all three prompts recover the time signature exactly. This is consistent with the self-consistency audit (Table 4), which also found the metadata fields to be deterministic. Metadata extraction is essentially model-bound rather than prompt-bound in this regime.

Second, on the *note-level* axis **Minimal** is statistically on par with **Original** on this 10-page sample (pitch-class F1 0.945 vs. 0.931; SER 0.135 vs. 0.156). This is initially counter-

intuitive because the minimal prompt does *not* explicitly tell the model how to write Jianpu (octave dots, dashes, trailing dots), yet for this VLM the model still produces a valid Jianpu note string. This suggests that for Gemini-2.5-flash on this corpus, the Jianpu-conventions half of the Appendix A prompt is carrying less of the load than the *schema* half. The Appendix A prompt buys robustness and predictability (same field names every time, plus a stable melody-role label; see also the observation on format drift below) rather than additional raw accuracy on the first line.

Third, **Chain-of-thought** is clearly worse: pitch-class F1 drops to 0.706 and SER rises to 0.413, even though all 10 pages returned parseable JSON. Inspection of the raw outputs shows that this is not because the describe-first step misreads the page—the natural-language description is usually correct—but because the follow-up JSON step drifts to an *inconsistent symbolic format*. Concretely, in the 10 CoT runs we observed (i) absolute Western pitch names (`A#4`, `D#5`) instead of Jianpu numerals, (ii) octave marks written as trailing apostrophes (`1'`) or as the letter `i` instead of as `^1`, and (iii) structured note objects of the form `{"pitch": 1, "octave": 1}` instead of the string field expected by the schema. These are all *plausible* representations of a pitch, but they are not the representation the downstream merge and QC stages (Section 3.7) are written against. CoT, therefore, degrades the pipeline not by damaging the model's perception of the score, but by destabilising the interface between the VLM and the rest of the system.

Two practical consequences follow for the claim this paper is making. First, the alt-prompt comparison demonstrates that the Jianpu-branch accuracy in Table 2 is not a fragile artefact of one particular prompt: a much simpler prompt reaches the same note-level F1 on the same pages with the same VLM. Second, a large share of the practical value of the Appendix A prompt is that it *fixes the symbolic format* the pipeline has to consume downstream; abandoning that format (as CoT does) does not help accuracy and actively hurts the end-to-end usability of the extracted material. This is consistent with Reviewer 3's framing of the paper as a systems-integration contribution: the prompt is part of the interface specification, not part of a novel recognition algorithm. A full comparison against specialised Jianpu-OMR systems such as those of [13,27,28], which would require re-implementing their preprocessing and rule bases, is out of scope for this revision and is listed explicitly as future work in Section 5.

5. Discussion

The prototype evaluation confirms that the MSMP framework successfully unifies heterogeneous music sources into a standardized symbolic representation. By integrating notation-aware VLM transcription with traditional symbolic parsing, the pipeline effectively bridges the gap between dispersed score repositories and computational musicology. The observed distribution of musical attributes—including key signatures, interval transitions, and rhythmic patterns—aligns with expected characteristics of traditional-style Chinese melody, suggesting that the extraction process preserves essential stylistic features despite the complexity of the input data.

However, several limitations constrain the current implementation. First, the reliance on proprietary API-based VLMs introduces latency and reproducibility challenges that a locally deployed model would mitigate. Second, the 50-page ground-truth benchmark reported in Section 4.7 is deliberately small and deliberately tilted toward printed metadata and first-line melody; it quantifies the dominant failure modes (accidental dropping, 2/4→4/4 collapse, staff pages being mis-routed to the Jianpu branch at the classifier stage) but does not yet provide a statistically tight estimate of note-level accuracy across an entire page or an entire corpus. Third, the structural QC gate (Section 3.7) does not attempt to detect content-level errors and, by design, will admit outputs that look plausible but drop

an accidental or collapse a meter; the ground-truth benchmark, not the QC gate, is the correct instrument for measuring such errors. Finally, the current monophonic assumption simplifies the processing of multi-voice scores, potentially merging accompaniment lines into the melody stream in complex grand-staff layouts. These factors identify the present work as a validation of the methodological framework and a first quantitative accuracy estimate on a small annotated subset, rather than the release of a definitive, large-scale benchmark corpus.

The most direct consequence of the ground-truth benchmark for future revisions of the pipeline is that content-level quality control cannot be delegated to the existing structural gate. Two complementary mechanisms are, therefore, planned for the next iteration. First, a *content-level QC layer* that applies targeted, inexpensive checks on top of the existing structural gate: cross-checking the declared key signature against the empirical pitch-class histogram of the transcribed melody (inconsistent pages are the ones most likely to have dropped an accidental), cross-checking the declared time signature against the number and rhythmic placement of beats per measure, and flagging pages where the classifier and the transcriber disagree on the notation family (the current benchmark already contains two such pages, both of which should be re-routed to the Audiveris branch rather than the VLM branch). The 100% inter-call agreement on key/meter/tempo observed in the self-consistency audit (Table 4) further suggests that a majority-vote over repeated calls will *not* rescue the accidental-dropping errors seen on the ground-truth benchmark—the errors are systematic, not stochastic, and so a *content-level* check is strictly required. Second, an *extended accuracy audit* in which the Jianpu ground-truth set is grown from 50 to a few hundred pages with at least two annotators per page so that Symbol Error Rate and Pitch-class F1 can be reported with confidence intervals rather than point estimates, and in which the 10-page self-consistency audit reported in Section 4.7 is itself scaled up to the full ground-truth set with a larger K, so that per-call variance can be separated from content-level error at every data point. We treat both of these items as first-class future work, not merely as suggested directions.

Further Systems-Level Limitations

Beyond the structural-vs-content QC gap, several systems-level limitations of the present prototype are worth stating explicitly, because they circumscribe what a reader can reasonably infer from the numbers reported in Sections 4.3 and 4.7.

Mixed pages are a declared, not yet fully implemented, branch. The score-type classifier is specified to emit `jianpu`, `staff`, `mixed`, and an invalid category, but the downstream treatment of the `mixed` label in the present prototype is intentionally conservative: `mixed` pages fall back onto the Jianpu VLM branch (on the assumption that a numeric row is always present), while a dedicated `mixed` handling path—for instance, cropping the Jianpu lines out of a grand-staff arrangement before transcribing—is still to be built. In practice this means the pipeline is currently well-characterised only on pages where a single notation family dominates; pages where the Jianpu numerals carry text and the staff lines carry the "real" melody (which are common in choral and vocal-plus-accompaniment arrangements) are an out-of-scope failure mode that we do not attempt to measure in this submission.

Melody-role assignment is not separately benchmarked. The Jianpu transcription prompt (Appendix A) instructs the model to "distinguish melody from lyrics or accompaniment when possible," and the structural QC gate requires at least one track labelled `melody`. We do not, however, measure the accuracy of this role label in isolation from the rest of the transcription, because the 50-page ground-truth set annotates only the melody line that a human reader would also identify; pages on which the VLM labels an accompaniment

row as `melody` would, therefore, show up in Section 4.7 as pitch-level errors rather than as role-assignment errors, and the two cannot be separated without a ground-truth labelling of *every* row on the page, which is not yet available. A targeted role-assignment benchmark is one of the specific items we list under "extended accuracy audit" above.

The monophonic simplification is a first-order limitation, not a peripheral one. The event schema is monophonic by construction, and the structural QC layer does not verify that an accepted page actually came from a monophonic source. In complex grand-staff or choral layouts, this means accompaniment lines may be merged into, or silently replace, the melody stream. This affects the validity of the extracted corpus for any downstream analysis that treats the melody as a single coherent voice; we report it here as a *first-order* limitation rather than a peripheral caveat, because any corpus-level statistic (including those in Section 4.6) would be distorted by even a small fraction of such merges.

Cultural and modal flattening. The schema maps Jianpu digits into absolute MIDI-equivalent pitches under a Western spelling convention (F major, G major, E major, etc.). This is computationally convenient—and is what makes the key/pitch metrics in Section 4.7 well-defined—but it over-simplifies the musical object it is describing. In particular, the current schema does not capture ornamentation marks, modal inflections (e.g., characteristic pentatonic degrees and their local alterations), or expressive glides that are common in traditional-style Chinese repertoire, and the Western-style key bins reported in Section 4.6 should be read as a *pitch-centroid summary* rather than as a modal analysis. We flag this as a methodological scope limitation and do not claim in this paper that the extracted corpus is adequate for fine-grained modal or ornamental analysis.

Collection dynamics are system-behaviour observations, not repertoire-level statistics. The yield and early-stopping numbers reported in Section 4.3 and the collection-dynamics paragraph of Section 4.2 are properties of one prototype run against two specific public portals under specific network availability. They are intended as documentation of the operational behaviour of the pipeline, not as stable properties of the underlying repertoire; in particular, the early-stopping behaviour is a function of source saturation and transient network availability rather than of an intrinsic size of the traditional-style Chinese melody corpus.

Reproducibility is partial, not total. The pipeline *logic*—routing, OMR back-end invocation, schema normalisation, structural QC, ground-truth evaluation, and self-consistency audit—is reproducible from the released source and the prompts in Appendix A. The *outputs*, however, depend on (i) unstable public portals that may change or disappear, (ii) a proprietary VLM whose behaviour is subject to provider-side versioning, and (iii) a manually annotated benchmark subset that is kept in the repository but is not yet accompanied by a released image set (the raw images are not redistributed for copyright reasons; see Section 6). We, therefore, describe MSMP, throughout this revision, as a *reproducible integration architecture* with a *partially reproducible* empirical run, and reserve the phrase "fully reproducible dataset release" for the future-work item listed in Section 6.

6. Conclusions

This paper presents MSMP, an automated multi-source pipeline for extracting traditional-style Chinese melody data from heterogeneous symbolic and score-image inputs. By unifying MIDI, MusicXML, Jianpu images, and staff images within one processing flow, and by combining notation-aware routing, a common event schema, and melody-oriented QC, MSMP turns a fragmented archival problem into a *partially reproducible* data-engineering procedure: the integration logic can be re-run from released code and prompts, while the empirical outputs of any given crawl remain dependent on public portals and a proprietary VLM (Section 5).

The main contribution of the study is methodological. The prototype case study demonstrates that the proposed extraction pipeline can produce structurally coherent symbolic outputs whose tonal, intervallic, rhythmic, and registral properties remain interpretable under musicological inspection. These observations support the feasibility of using MSMP as a front-end for future corpus construction in this repertoire. The modular architecture, incremental processing strategy, and VLM-assisted Jianpu transcription together provide a practical foundation for future data collection, evaluation, and downstream symbolic music research.

We did *not* bundle the corpus produced during the prototype run with this submission. This deliberate choice is driven by two concerns that we believe a corpus release deserves before, rather than after, peer review. First, the raw inputs are retrieved from third-party score portals and, therefore, inherit an unresolved mixture of copyright statuses; redistributing a large extracted corpus without clearing these rights piece by piece would create exactly the kind of legal ambiguity that responsible data releases should avoid. Second, we have begun the accuracy audit in this revision, but only on a 50-page manually annotated subset (Section 4.7); the full-scale audit, with several hundred pages and multiple annotators, is part of the planned future work (Section 5). Publishing a large corpus before that full audit would risk overstating the quality of the release. To still make the methodology concrete and inspectable, we plan to release, alongside the next version of the pipeline, a small set of derived per-segment statistics and schema-only examples, together with the pipeline code, so that independent groups can regenerate comparable corpora from their own legally cleared inputs.

Future work will focus on five key directions to address current limitations. First, replacing the proprietary VLM with a fine-tuned local model (e.g., Qwen-VL [34], InternVL [35]) to improve throughput, privacy, and reproducibility. Second, scaling the manually annotated Jianpu benchmark from 50 pages to several hundred pages with multiple annotators, so that the headline accuracy numbers reported in Section 4.7 can be reported with confidence intervals rather than point estimates, and so that a targeted sub-benchmark of staff pages mis-routed to the Jianpu branch can be used to retrain the notation-aware classifier. Third, introducing a content-level QC layer that uses inexpensive musicological cross-checks (key-vs-histogram consistency, meter-vs-beat-count consistency) on top of the existing structural QC gate, together with an expanded self-consistency protocol that scales the 10-page, $K = 3$ audit of Section 4.7 to the full ground-truth set with a larger number of repeated calls per page, so that plausible-looking transcriptions that silently drop accidentals or collapse a 2/4 m can be flagged before they reach the accepted corpus. Fourth, extending the pipeline to handle polyphonic textures and lyrics alignment, thereby broadening the scope of the extracted data for multi-modal research tasks. Fifth, and most directly motivated by the present reviewer round, preparing a publicly releasable, copyright-cleared, accuracy-audited traditional-style Chinese melody corpus that the present pipeline can produce end-to-end.

Author Contributions: Conceptualization, X.Z., Y.H., S.H. and J.B.; methodology, X.Z., Y.H., S.H. and J.B.; software, X.Z., Y.H., S.H. and J.B.; validation, X.Z., Y.H., S.H. and J.B.; data curation, X.Z., Y.H., S.H. and J.B.; writing—original draft preparation, X.Z., Y.H., S.H. and J.B.; writing—review and editing, X.Z., Y.H., S.H. and J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The pipeline source code, manuscript assets, and derived per-segment statistics used to generate the figures in this study are available from the authors upon reasonable request. Raw score images retrieved from third-party portals are *not* redistributed, because their

individual copyright statuses have not been cleared; each input page in our logs is traced back to its public source URL so that independent researchers can regenerate comparable corpora from their own legally cleared inputs. A release of the extracted symbolic corpus is planned once a copyright audit and a note-level accuracy audit against manually annotated Jianpu references have been completed (see Section 6).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Model Configuration and Prompt Templates

This appendix records the exact model, inference parameters, and prompts used in the VLM branch so that the reported behaviour can be reproduced independently of the specific access channel.

Appendix A.1. Model Access and Inference Parameters

Both the score-type classifier and the Jianpu transcriber are invoked through a single OpenAI-compatible chat-completions endpoint that wraps *gemini-2.5-flash* [32]. For this reason, the pipeline does not use Google's native Python SDK or a `GenerationConfig` object directly; instead, the parameters below are passed through the OpenAI-compatible interface and then mapped by the provider to the corresponding Gemini controls.

Table A1. Inference configuration used for all VLM calls in the prototype.

Parameter	Value Used	Nearest Gemini `GenerationConfig` Field
`model`	`gemini-2.5-flash`	`model`
`system_message`	per-task system prompt (see Appendix A.2/Appendix A.3); played the role of Gemini's `system_instruction`	`system_instruction`
`tools`	none (no function calls; no Google Search grounding)	`tools`
`temperature`	0.7	`temperature`
`top_p`	not set; provider default applied	`top_p`
`top_k`	not exposed by the OpenAI-compatible wrapper	`top_k`
`max_tokens`	65,536	`max_output_tokens`
`response_format`	strict JSON required by the prompt; malformed output is discarded, see Section 3.5	`response_mime_type/response_schema`
`stop`	not set	`stop_sequences`
`safety_settings`	provider default (not overridden by the client)	`safety_settings`
`random_seed`	not set at the API level	`seed` (where supported)
`timeout`	180 s per request	(transport)
`retries/retry_delay`	10/20 s	(client)
image input	single-turn; image embedded as a base64 data URL alongside the text prompt	vision-input part

For the score-type classifier, the same model and decoding configuration are reused; the operational difference with respect to the Jianpu transcriber lies only in the system message and expected output schema (Appendixes A.2 and A.3 below). Because decoding is performed with a non-zero `temperature`, individual outputs are not bit-identical across reruns; however, the acceptance/rejection behaviour of the melody-oriented QC

layer is deterministic given an output, and all inputs are deduplicated by SHA-256 hash before inference.

Appendix A.2. Score-Type Classifier Prompt

System message (played the role of Gemini's `system_instruction`):

```
You are a score-classification assistant.
```

User message (text part; the image is sent alongside):

```
Analyze the input score image and return JSON only:
{
  "type": "staff" | "jianpu" | "mixed" | "lyrics_only" | "junk",
  "layout": "single_staff" | "grand_staff" | "score" | "unknown",
  "quality": "high" | "low" | "incomplete",
  "content": {
    "has_lyrics": boolean,
    "has_chords": boolean
  }
}
```

Appendix A.3. Jianpu Transcription Prompt

System message (played the role of Gemini's `system_instruction`):

```
You are a Jianpu transcription expert.
```

User message (text part; the image is sent alongside):

```
Convert the input score image into structured JSON data.

Rules:
1. Identify the key declaration (for example 1=D) and time signature.
2. Read notes in musical order and encode pitch with digits 1-7.
   Use 0 for rests.
3. Use ^ for upper-octave dots and _ for lower-octave dots.
4. Encode duration in beats:
   - bare digit = quarter note
   - underline = halve the duration
   - dash = extend by one beat
   - dot = extend by one half
5. Distinguish melody from lyrics or accompaniment when possible.

Return JSON in the following form:
{
  "key": "D",
  "time_signature": "4/4",
  "bpm": 90,
  "parts": [
    {
      "role": "melody",
      "measures": [
        [ {"note": "1", "duration": 1.0},
```

```
                {"note": "2", "duration": 1.0},
                {"note": "3", "duration": 2.0} ]
            ]
        }
    ]
}
```


References

1. Deng, J.; Tang, Y. Music Information Retrieval in the Deep Learning Era: A Comprehensive Review. *Expert Syst. Appl.* **2024**, *240*, 122565. [CrossRef]
2. Ji, S.; Luo, J.; Yang, X. A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions. *arXiv* **2020**, arXiv:2011.06801. [CrossRef]
3. Ma, Y.; Øland, A.; Ragni, A.; Del Sette, B.M.; Saitis, C.; Donahue, C.; Lin, C.; Plachouras, C.; Benetos, E.; Shatri, E.; et al. Foundation Models for Music: A Survey. *arXiv* **2024**, arXiv:2408.14340. [CrossRef]
4. Müller, M. *Fundamentals of Music Processing*; Springer: Cham, Switzerland, 2015; ISBN 978-3-319-21944-8.
5. Schedl, M.; Gómez, E.; Urbano, J. Music Information Retrieval: Recent Developments and Applications. *Found. Trends Inf. Retr.* **2014**, *8*, 127–261. [CrossRef]
6. Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; Défossez, A. Simple and Controllable Music Generation (MusicGen). In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023.
7. Huang, C.-Z.A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Simon, I.; Hawthorne, C.; Dai, A.M.; Hoffman, M.D.; Dinculescu, M.; Eck, D. Music Transformer: Generating Music with Long-Term Structure. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
8. Raffel, C. Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching. Ph.D. Thesis, Columbia University, New York, NY, USA, 2016. Available online: https://colinraffel.com/publications/thesis.pdf (accessed on 1 March 2026).
9. Hawthorne, C.; Stasyuk, A.; Roberts, A.; Simon, I.; Huang, C.-Z.A.; Dieleman, S.; Eck, D. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019. Available online: https://openreview.net/forum?id=r1lYRjC9F7 (accessed on 1 March 2026).
10. Du, Y. *Introduction to Chinese National Music (Zhongguo Minzu Yinyue Gailun)*, 2nd ed.; Shanghai Music Publishing House: Shanghai, China, 2002; ISBN 978-7805530348.
11. Zhang, Y. Eastern Rhythmic Foot and Western Colors: The Cross-Cultural Practice of Chinese Pentatonic Scales in Impressionist Music. *J. Lit. Arts Res.* **2025**, *2*, 1–11.
12. Calvo-Zaragoza, J.; Hajič Jr., J.; Pacha, A. Understanding Optical Music Recognition. *ACM Comput. Surv.* **2020**, *53*, 1–35. [CrossRef]
13. Wu, F.-H. Applying Machine Learning in Optical Music Recognition of Numbered Music Notation. *Int. J. Multimed. Data Eng. Manag.* **2017**, *8*, 21–41. [CrossRef]
14. Li, S.; Wu, Y. An Introduction to a Symbolic Music Dataset of Chinese Guqin Pieces and Its Application Example. *J. Fudan Univ. (Nat. Sci.)* **2020**, *59*, 276–285.
15. Cuthbert, M.S.; Ariza, C. Music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Utrecht, The Netherlands, 9–13 August 2010; pp. 637–642.
16. Bitteur, H. *Audiveris: An Open-Source OMR Engine*, version 5.3; 2023. Available online: https://audiveris.github.io/audiveris/ (accessed on 1 March 2026).
17. Kong, Q.; Li, B.; Chen, J.; Wang, Y. GiantMIDI-Piano: A Large-Scale MIDI Dataset for Classical Piano Music. *arXiv* **2020**, arXiv:2010.07061. [CrossRef]
18. Donahue, C.; Mao, H.H.; Li, Y.E.; Cottrell, G.W.; McAuley, J. LakhNES: Improving Multi-Instrumental Music Generation with Cross-Domain Pre-Training. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Delft, The Netherlands, 4–8 November 2019; pp. 685–692.
19. Wang, Z.; Chen, K.; Jiang, J.; Zhang, Y.; Xu, M.; Dai, S.; Bin, G.; Xia, G. POP909: A Pop-song Dataset for Music Arrangement Generation. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Montreal, QC, Canada, 11–15 October 2020; pp. 38–45.

20. Pfleiderer, M.; Frieler, K.; Abeßer, J.; Zaddach, W.-G.; Burkhart, B. (Eds.) *Inside the Jazzomat: New Perspectives for Jazz Research*; Schott Campus: Mainz, Germany, 2017. Weimar Jazz Database. Available online: https://jazzomat.hfm-weimar.de/dbformat/dboverview.html (accessed on 20 March 2026).
21. van Kranenburg, P.; de Bruin, M.; Grijp, L.; Wiering, F. The Meertens Tune Collections: The Annotated Corpus (MTC-ANN) Versions 1.1 and 2.0.1. Meertens Online Reports, 2016. Available online: https://www.liederenbank.nl/mtc/ (accessed on 20 March 2026).
22. Simonetta, F.; Carnovalini, F.; Orio, N.; Rodà, A. Symbolic Music Similarity through a Graph-Based Representation. In Proceedings of the Audio Mostly Conference, Nottingham, UK, 18–20 September 2018.
23. Gotham, M.; Jonas, P.; Bower, B.; Bosma, W.; Bergomi, M.; Couturier, L.; Dang, L. Scores of Scores: An öMNES Opus and its Community-Driven Curation. In Proceedings of the International Conference on Digital Libraries for Musicology (DLfM), Budapest, Hungary, 28 July 2018; pp. 87–95. The Josquin Research Project. Available online: https://josquin.stanford.edu/ (accessed on 20 March 2026).
24. Thickstun, J.; Harchaoui, Z.; Kakade, S.M. Learning Features of Music from Scratch. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017. Available online: https://openreview.net/forum?id=rkFBJv9gg (accessed on 9 March 2026).
25. Zhou, M.; Xu, S.; Liu, Z.; Wang, Z.; Yu, F.; Li, W.; Han, B. CCMusic: An Open and Diverse Database for Chinese Music Information Retrieval Research. *Trans. Int. Soc. Music Inf. Retr.* **2025**, *8*, 22–38. Available online: https://transactions.ismir.net/articles/10.5334/tismir.194 (accessed on 9 March 2026). [CrossRef]
26. Baró, A.; Riba, P.; Calvo-Zaragoza, J.; Fornés, A. From Optical Music Recognition to Handwritten Music Recognition: A Baseline. *Pattern Recognit. Lett.* **2019**, *123*, 1–8. [CrossRef]
27. Krishnan, R.; Natarajan, B.; Vadivel, M. Numbered Musical Notation Recognition via Deep Layout Analysis and Template Matching. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), San José, CA, USA, 21–26 August 2023.
28. Bu, F.; Li, R.; Li, Z.; Li, Y. The Renaissance of Expert Systems: Optical Recognition of Printed Chinese Jianpu Musical Scores with Lyrics. *arXiv* **2025**, arXiv:2512.14758. [CrossRef]
29. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; Volume 36.
30. Bertin-Mahieux, T.; Ellis, D.P.W.; Whitman, B.; Lamere, P. The Million Song Dataset. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Miami, FL, USA, 24–28 October 2011; pp. 591–596.
31. Vigliensoni, G.; Burlet, G.; Fujinaga, I. Optical Measure Recognition in Common Music Notation. In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR), Curitiba, Brazil, 4–8 November 2013; pp. 125–130. Available online: http://ismir2013.ismir.net/wp-content/uploads/2013/09/207_Paper.pdf (accessed on 9 March 2026).
32. Google. Gemini 2.5 Flash [Software Documentation]. 2025. Available online: https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash (accessed on 9 March 2026).
33. Narmour, E. *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*; University of Chicago Press: Chicago, IL, USA, 1990; ISBN 978-0226568425.
34. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv* **2023**, arXiv:2308.12966.
35. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. InternVL: Scaling Up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024. Available online: https://github.com/OpenGVLab/InternVL (accessed on 1 March 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.