Zero-shot Jianzi Recognition as Structured Visual Information Extraction in Open Compositional Symbolic Systems

Zehan Li, Fu Zhang*, Zhijun Liu, Jingwei Cheng
School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
{lizehan1999,liuzhijun_2003}@163.com, {zhangfu,chengjingwei}@neu.edu.cn

Abstract

Guqin (古琴) Jianzi (減字) is an open and freely compositional tablature system that encodes performance actions rather than acoustic outcomes. Its automatic recognition remains largely unexplored, as conventional OCR assumes a closed and enumerable glyph set and struggles with Jianzi's unbounded composition and manuscript-level variability.

We introduce **Zero-shot Jianzi Recognition**, which formulates Jianzi recognition as vision-to-sequence prediction of canonical component sequences under a zero-shot split. To enable scalable supervision, we construct Synthetic-JZ from aligned online composition metadata. We then synthesize manuscript-like training images via component-wise style recomposition and manuscript-domain noise modeling, and fine-tune a VLM for end-to-end component sequence recognition. At inference time, a lightweight legality-guided correction module re-ranks decoding candidates, suppressing structural hallucinations without modifying the backbone.

Experiments on two benchmarks show that our method achieves 63.02% sequence accuracy on Real-JZ, our manually annotated real-world Jianzi benchmark, surpassing Gemini-3-Pro by 35.11%. This result highlights the feasibility of reliable automated Jianzi recognition and its potential for large-scale digitization of historical Guqin Jianzi Pu manuscripts.

Figure 1: **(a)** The spatial structure of the Guqin (古琴), showing its seven strings and thirteen hui (徽) positions. **(b)** An excerpt from a Jianzi Pu (減字譜) manuscript of *Flowing Water*, where the circled Jianzi glyph is shown as an example. **(c)** Visual variability of the same Jianzi glyph across different manuscripts. **(d)** Structural decomposition of a Jianzi glyph into semantic components derived from Chinese characters, which are spatially recombined to form a compact glyph encoding a complete performance instruction.

1 Introduction

Optical Character Recognition (OCR) has made substantial progress in recent years (Wang et al., 2023; Yang et al., 2025b), and large Vision–Language Models (VLMs) have shown strong zero-shot ability (Tong et al., 2025; Xu et al., 2024) for general visual understanding and text-related reasoning (Semnani et al., 2025). Nevertheless, both paradigms are most reliable under standardized and structurally stable symbol systems: OCR typically assumes a closed and enumerable label space, while VLMs rely on pretrained priors and may generate visually plausible yet structurally inconsistent outputs when faced with unfamiliar symbol composition (Zhang et al., 2025a). For open, highly compositional, and visually variable notation, these assumptions no longer hold, and recognition failures often manifest as struc-

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29414–29429

tural hallucinations (Simon et al., 2025).

The Guqin (古琴) Jianzi (減字) tablature system is a representative yet long-overlooked example of such an open symbolic system. Guqin (Figure 1(a)) is a traditional Chinese plucked zither with over three millennia of history and was inscribed as *UNESCO Intangible Cultural Heritage* in 2003. Its primary written medium, Jianzi Pu (減字譜) (Figure 1(b)), encodes performance actions, including fingering techniques, string indices, hui (徽) positions, and expressive markers, rather than acoustic outcomes. Each Jianzi glyph spatially recombines multiple semantic components into a compact character-like form, yielding a hybrid notation system distinct from both staff notation and logographic writing (Lee, 2023).

Despite its visual resemblance to Chinese characters, Jianzi differs fundamentally from writing systems in **three key aspects**. ***First***, its appearance is highly variable: the same glyph may differ substantially across manuscripts, print editions, calligraphic styles, and transmission lineages (Figure 1(c)). ***Second***, its structure is open and compositional: a glyph consists of a flexible number of semantic components arranged under loose spatial constraints, producing an effectively unbounded set of composite symbols (Figure 1(d)). ***Third***, Jianzi recognition is extremely low-resource: most surviving sources are handwritten or degraded, no unified encoding standard exists, and annotation requires expert-level domain knowledge (Kuremoto et al., 2025). Together, these properties make closed-vocabulary OCR formulations ineffective for scalable Jianzi digitization.

Existing attempts at Jianzi recognition primarily adopt CNN-based glyph classification trained on small manually annotated datasets (Wei and Wang, 2023; Hayami et al., 2025). Such methods implicitly assume a fixed symbol inventory and therefore struggle to generalize beyond seen glyphs. We argue that Jianzi recognition should instead be treated as vision-to-sequence transcription of canonical component sequences, i.e., **structured visual information extraction** that parses internal components and preserves their functional ordering. This formulation better matches the compositional nature of Jianzi Pu and provides a concrete testbed for probing how modern VLMs handle open compositional symbolic systems beyond closed character inventories.

In this work, we introduce **Zero-shot Jianzi Recognition**, which maps an image of a segmented Jianzi glyph to its canonical component sequence under a zero-shot setting, where no complete sequence is shared between training and test. A key challenge lies in supervision: large-scale manual annotation is costly, while purely synthetic glyphs lack manuscript realism. To address this, we construct Synthetic-JZ from online composition metadata, retaining only samples with exact component−metadata alignment. We further reduce the manuscript domain gap by synthesizing manuscript-like images via **component-wise style recomposition** and **manuscript noise modeling**. Using this augmented corpus, we finetune a VLM for end-to-end component sequence prediction. At inference time, we introduce a lightweight **legality-guided correction module** that re-ranks decoding candidates using a structural prior, selecting the most consistent sequence without modifying the backbone.

Our contributions are fourfold: (1) We introduce **Zero-shot Jianzi Recognition (ZJR)**, formulating Jianzi transcription as structured visual information extraction instead of a conventional classification problem. (2) We construct **Synthetic-JZ**, a large-scale corpus derived from online composition metadata, enabling controlled vision-to-sequence training. (3) We propose **a ZJR framework named JZ-Glyph**, which integrates component-wise style recomposition, manuscript noise modeling, and a legality-guided correction module for inference-time structural correction. (4) We evaluate under zero-shot and cross-domain settings; on the manually annotated **Real-JZ** benchmark from historical manuscripts, our method achieves 63.02% accuracy, outperforming Gemini-3-Pro by 35.11%.

2 Related Work

2.1 Guqin Jianzi Recognition

Recent studies on the Guqin have explored audio analysis (Huang et al., 2020), knowledge graph construction (Zhou et al., 2025), and cultural dissemination (Yu et al., 2021). However, Jianzi Pu (減字譜), the primary tablature system for Guqin performance, remains a major obstacle to digital preservation due to its highly compressed and compositional structure. Although more than 3,000 tablature books survive, only a small fraction have been transcribed (Shi et al., 2024), as interpretation requires expert-level knowledge.

Early Jianzi recognition relied on heuristic de-

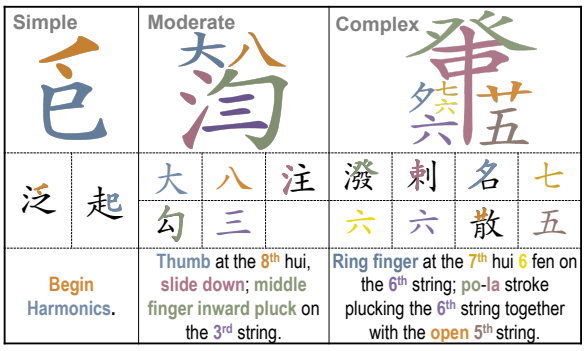

Figure 2: Representative Jianzi symbols grouped by structural complexity defined by the number of components: Simple (1–3), Moderate (4–6), and Complex (7+), with corresponding performance instructions.

composition (Ni et al., 2010), later replaced by CNN-based classification methods (Shi, 2016; Repolusk and Veas, 2025). Some works collect handwritten samples and apply data augmentation (Yang et al., 2023; Kuremoto et al., 2025), while recent radical-based approaches introduce limited structural flexibility (Hayami et al., 2025). However, these methods are typically developed and evaluated on relatively small, closed, and often non-public datasets. They formulate Jianzi recognition as a closed-set visual classification problem that treats each symbol as an isolated category rather than a structured instruction. Consequently, their formulations and evaluation settings differ from our zero-shot, open-set setting, limiting the direct comparability of results.

To overcome these limitations, we formulate Jianzi recognition as a visual information extraction problem. Unlike optical music recognition for staff notation (Tang et al., 2025), Jianzi recognition requires parsing open-ended component compositions and decoding their semantic action sequences. Vision–Language Models (VLMs) (Yang et al., 2025a) naturally support end-to-end image-to-sequence generation, enabling zero-shot generalization to unseen Jianzi.

2.2 Chinese Character Recognition

Chinese Character Recognition (CCR) focuses on mapping glyph images to predefined character categories (Dai et al., 2007). To address vocabulary growth and long-tail distributions, recent work explores zero-shot recognition of Out-Of-Vocabulary characters (Zhang et al., 2025b). Many approaches exploit character decomposability by aligning glyphs with IDS or component representations (Yu et al., 2023; Zhang et al., 2025c), or by synthesizing unseen samples for classifier calibration (Ao et al., 2025). CCR techniques have also been extended to historical scripts and calligraphy (Guan et al., 2024; Bao et al., 2025a,b).

However, these frameworks are not directly applicable to Jianzi, which is an open, freely compositional notation system with extreme data sparsity and no fixed symbol inventory. Existing zero-shot CCR methods typically assume stable glyph structures or predefined decomposition grammars, assumptions that do not hold for Jianzi.

In contrast, we address Jianzi recognition in an extremely low-resource and open-set setting by combining component-aligned data construction, manuscript-style synthesis, and vision-to-sequence modeling, together with a lightweight legality-guided correction module, to enable effective and structurally consistent Jianzi recognition.

3 Methodology

3.1 Task Formulation

We formulate zero-shot Jianzi recognition as an end-to-end sequence prediction problem. Given an input image x of a single segmented Jianzi glyph, the model predicts a component sequence $y = (v_1, \ldots, v_T)$ with components $v_t \in \mathcal{V}$, where $\mathcal{V}$ is a vocabulary of canonical Jianzi components, each encoding a specific performance primitive such as technique, timbre, fingering, string index, hui position, or expressive marker. The output space is $\mathcal{Y} = \bigcup_{T \geq 1} \mathcal{V}^T$. Under the zero-shot setting, no complete component sequence appears in both training and test sets, i.e., $\mathcal{Y}_{\text{train}} \cap \mathcal{Y}_{\text{test}} = \varnothing$. As an open compositional system, Jianzi exhibits varying structural complexity defined by the number of semantic components (Figure 2).

3.2 Overview

Figure 3 illustrates the overall pipeline. We first construct a large-scale Jianzi dataset of aligned image–label pairs from online composition metadata (§3.3). Based on these samples, we synthesize manuscript-like training data via **Compositional Style Simulation** (§3.4), including **Component-level Style Recomposition** (§3.4.1) and **Manuscript-domain Noise Modeling** (§3.4.2). The resulting data are used to train a VLM for canonical component sequence prediction, with a lightweight **Legality-guided Correction Module** (§3.5) applied at inference time to

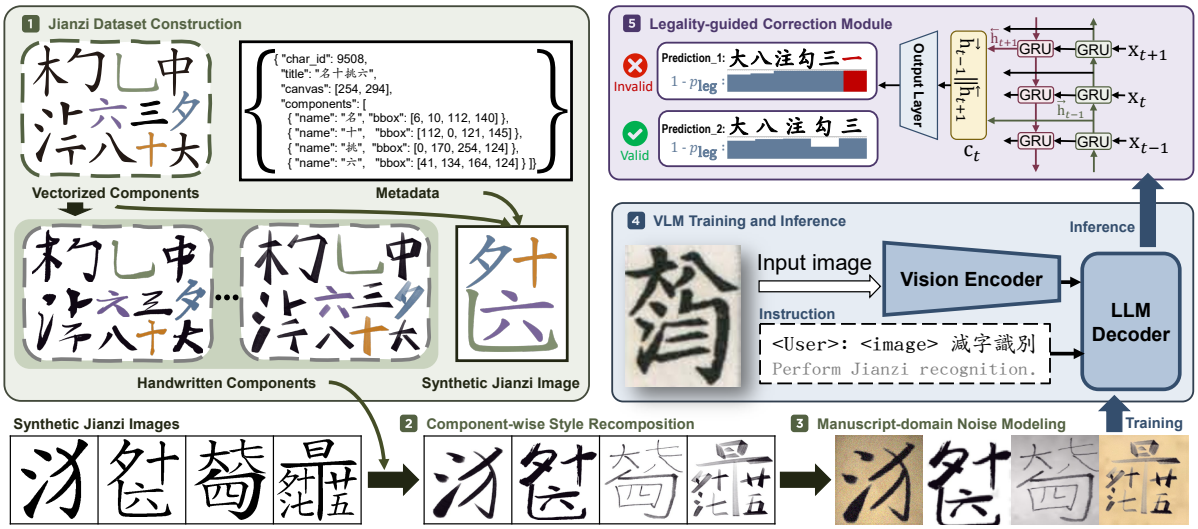

Figure 3: Overview of the proposed **JZ-Glyph** framework. **(1)** Large-scale **Jianzi dataset construction** from vectorized components and metadata. **(2) Component-wise style recomposition** using handwritten variants to synthesize diverse glyphs. **(3) Manuscript-domain noise modeling**, simulating real Jianzi Pu capture and degradation effects. **(4) VLM training and inference** for mapping Jianzi images to canonical component sequences. **(5)** A lightweight **legality-guided correction module** for filtering structurally invalid predictions.

suppress structurally invalid outputs.

3.3 Synthetic-JZ Dataset Construction

Existing research on Jianzi recognition faces three challenges: (1) the absence of any publicly available large-scale annotated dataset; (2) the mismatch between closed-vocabulary OCR assumptions and Jianzi's open, compositional structure; and (3) the high cost of manual annotation due to its reliance on expert knowledge. To address these challenges, we construct a large-scale synthetic dataset, termed Synthetic-JZ, from composition records of a Guqin community platform[1].

The platform allows practitioners to compose Jianzi symbols by assembling predefined vectorized components and assigning textual labels that roughly correspond to canonical component combinations. We collect **fourteen years** of such construction records, which encode extensive expert knowledge accumulated from real-world Guqin tablature practice. While these records exhibit substantial structural diversity and reflect authentic compositional patterns, they also contain label noise, inconsistency, and redundancy introduced by free-form user input.

To obtain a structurally reliable corpus, we perform systematic data cleaning and canonicalization by decomposing each user-provided label into its constituent components and retaining only samples with exact component–metadata alignment (where metadata specifies the component type, relative position, and scale; see Figure 3 (1)). Structurally complex or ambiguous cases are further inspected and manually verified. The resulting Synthetic-JZ dataset provides large-scale, component-aligned Jianzi samples with weak semantic supervision, preserving the open and compositional nature of Jianzi while enabling controlled vision-to-sequence training. Dataset statistics are reported in Appendix A.

3.4 Compositional Style Simulation

While our Synthetic-JZ dataset precisely encodes structural composition, its visual appearance is limited to a uniform digital template style, lacking the variability of handwritten or printed Jianzi Pu manuscripts and thus constraining real-world generalization. To bridge this gap, we introduce a compositional style simulation pipeline that injects realistic appearance variation.

3.4.1 Component-wise Style Recomposition

A common approach for bridging the gap between synthetic and handwritten data is character-level font or style generation (Yang et al., 2024). However, such methods *almost completely fail* on Jianzi, whose open and freely compositional structure violates the fixed topological assumptions underlying character-level style mapping.

To address this limitation, we shift from

[1] https://www.guanglingsan.com/

character-level to **component-level** style transfer. We collect handwritten variants of high-frequency Jianzi components from Chinese volunteers using tablets, and recompose full glyphs by replacing vectorized components according to their metadata, as illustrated in Figure 3 (2).

This approach injects authentic component-level variation while preserving strict structural control, thereby narrowing the domain gap between synthetic and real data. In addition, it avoids the prohibitive cost of annotating complete Jianzi glyphs, since a small inventory of frequent components can be recombined to synthesize large-scale, stylistically diverse samples.

3.4.2 Manuscript-domain Noise Modeling

Despite stylistic recomposition, synthetic Jianzi still lack the imaging artifacts commonly observed in real Jianzi Pu manuscripts. To narrow this gap, we introduce a unified noise injection pipeline that simulates realistic manuscript capture and degradation effects. Specifically, we apply geometric perturbations to model viewpoint variation and paper deformation, together with resolution degradation and Gaussian noise to mimic low-quality scanning. We further simulate manuscript-specific disturbances by shifting or cropping page margins and compositing Jianzi glyphs onto textured or aged paper backgrounds.

This process preserves the recognizability of Jianzi structures while introducing realistic-scale variations. The resulting augmented data better matches the visual distribution of real Jianzi Pu sources and is used to train the VLM, improving performance in real-world settings.

3.5 Legality-guided Correction Module

Although the VLM trained in previous stages can map Jianzi images to component sequences, its generative nature makes it prone to structural hallucinations, i.e., outputs that are visually plausible yet violate the syntactic constraints of Jianzi notation (e.g., a single string action followed by two string numbers). To address this issue without modifying the VLM architecture, we introduce a lightweight legality-guided correction module based on a Bi-GRU structural prior trained solely on component sequences from the training set.

Structural prior. Given a Jianzi component sequence $y = (v_1, \ldots, v_T)$, a Bi-GRU is applied to model local bidirectional dependencies. For each position t, a context representation is computed as

$$\mathrm{c}_t = \overrightarrow{\mathrm{h}}_{t-1} \,\|\, \overleftarrow{\mathrm{h}}_{t+1}, \quad \overrightarrow{\mathrm{h}}_t, \overleftarrow{\mathrm{h}}_t = \mathrm{GRU}(\mathrm{x}_t), \quad (1)$$

where x_t denotes the embedding of component v_t. The conditional legality of a component $\hat{v}_t$ under its local structural context is then defined as

$$p_{\mathrm{leg}}(\hat{v}_t \mid \mathrm{c}_t) = \mathrm{Softmax}(\mathrm{W}_p \mathrm{c}_t + \mathrm{b})_{\hat{v}_t}. \quad (2)$$

Sequence-level legality. Structural validity is a sequence-level property rather than a local one. We therefore aggregate token-wise legality scores into a normalized sequence legality score

$$S_{\mathrm{leg}}(y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log p_{\mathrm{leg}}(v_t \mid \mathrm{c}_t), \quad (3)$$

and define an anomaly ratio, which measures the proportion of structurally illegal components:

$$R_{\mathrm{ill}}(y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \mathbb{I}\big[p_{\mathrm{leg}}(v_t \mid \mathrm{c}_t) < \tau\big]. \quad (4)$$

$\tau = 0.05$ is a fixed threshold empirically chosen to identify low-confidence structural components.

Legality-guided correction. We generate a candidate set $\mathcal{C}(x) = \{y^{(1)}, \ldots, y^{(K)}\}$ with $K = 5$ using beam-based decoding during inference. The legality module performs correction by selecting the most structurally consistent candidate:

$$y^* = \arg\min_{y \in \mathcal{C}(x)} \Big(R_{\mathrm{ill}}(y), -S_{\mathrm{leg}}(y)\Big), \quad (5)$$

where candidates are first ranked by their anomaly ratio and ties are broken by the sequence-level legality score. The selected output y^* thus suppresses structural hallucinations without modifying the VLM parameters.

4 Experimental Setup

4.1 Datasets and Benchmarks

Following §3.3, we evaluate Jianzi recognition using two benchmarks under a zero-shot setting. Given the open and compositional nature of Jianzi, we ensure that no component sequence is shared across training, validation, or test splits.

(1) Synthetic-JZ corresponds to the test split of the full Synthetic-JZ dataset (§3.3), and is used to evaluate performance in a controlled synthetic setting with diverse component compositions.

Model	Synthetic-JZ					Real-JZ				
	P. ↑	R. ↑	F1 ↑	CER ↓	Acc ↑	P. ↑	R. ↑	F1 ↑	CER ↓	Acc ↑
Frozen Parameters ❄										
GPT-5.1	41.39	34.06	35.47	82.17	9.30	22.57	21.15	21.03	58.36	15.86
Qwen-VL-Max	44.44	38.33	39.39	88.66	4.07	41.54	34.40	35.76	83.83	8.53
Qwen3-VL-Plus	60.88	48.31	52.85	58.55	6.40	43.62	33.67	36.58	76.05	9.30
Gemini-2.5-Flash	63.95	52.59	56.77	52.41	12.79	54.89	44.90	47.79	66.85	20.93
Gemini-3-Pro-Preview	72.14	59.98	64.60	42.66	17.44	66.69	56.08	59.08	54.43	27.91
Fine-tuned Parameters 🔥										
PARSeq (Bautista and Atienza, 2022)	89.67	88.99	89.11	15.11	61.69	41.62	39.76	40.05	68.82	13.77
CCR-CLIP (Yu et al., 2023)	82.90	82.93	82.21	26.06	67.48	42.78	42.16	39.68	94.85	13.89
TrOCR (Li et al., 2023)	96.02	95.77	95.76	7.25	82.48	72.19	72.58	71.98	36.34	45.58
YOLO-Radical (Hayami et al., 2025)	80.89	78.08	78.88	30.63	48.46	28.03	23.90	24.14	94.16	5.94
JZ-Glyph (ours)	**96.56**	**96.12**	**96.29**	**5.79**	**85.39**	**85.19**	**83.93**	**84.20**	**19.45**	**63.02**
*Backbone Variants of **JZ-Glyph***										
LLaVA-1.5-7B (Liu et al., 2023)	94.43	93.85	94.03	9.02	79.47	60.91	60.75	60.38	48.86	34.29
LLaVA-NeXT-7B (Li et al., 2024)	94.84	94.12	94.36	8.26	80.30	59.78	58.82	58.81	49.88	32.62
InternVL3-8B (Zhu et al., 2025)	96.12	95.64	95.81	6.43	83.31	74.14	74.14	73.75	33.99	48.78
InternVL3.5-8B (Wang et al., 2025)	96.04	95.75	95.85	6.28	84.53	79.71	79.31	79.20	25.85	55.11
Qwen2-VL-7B (Wang et al., 2024)	93.37	92.34	92.72	10.31	71.74	69.86	70.93	69.85	38.97	40.83
Qwen2.5-VL-7B (Bai et al., 2025)	94.63	93.83	94.11	9.19	77.98	73.01	72.03	72.00	34.67	46.00

Table 1: Main results on Synthetic-JZ and Real-JZ. JZ-Glyph denotes the proposed framework for zero-shot Jianzi recognition, with Qwen3-VL-8B as the default backbone. ***Frozen-parameter*** ❄ models are state-of-the-art VLMs used directly via official APIs. ***Fine-tuned parameter*** 🔥 models are trained on the Synthetic-JZ training set. *Backbone variants* show the performance of JZ-Glyph under different VLM backbones. Best results are in **bold**.

(2) Real-JZ is a manually annotated benchmark curated from historical Jianzi Pu manuscripts. All sequences in Real-JZ are disjoint from the training set, enabling evaluation of generalization from synthetic data to real-world manuscripts with handwriting variation, manuscript noise, and unseen compositional patterns.

Dataset statistics and details of the manual annotation process are reported in Appendix A, with representative manuscript examples provided in Appendix E. Code and data are available at https://github.com/lizehan1999/JZ-Glyph/.

4.2 Implementation Details

We implement **JZ-Glyph** with multiple VLM backbones (LLaVA, InternVL, Qwen), using **Qwen3-VL-8B** (Yang et al., 2025a) as the default. LoRA fine-tuning is applied to all backbones with a learning rate of 2e-4 for 5 epochs. The legality-guided correction module is trained with a 128-dimensional embedding, a learning rate of 1e-3, a batch size of 256, and 10 epochs. All experiments are conducted on four NVIDIA RTX 4090 GPUs.

For evaluation, we adopt standard OCR metrics: Sequence Accuracy (**ACC**) measures exact sequence matching, while Character Error Rate (**CER**) computes the edit-distance ratio between predictions and ground truth. To assess fine-grained recognition quality, we further report character-level Precision (**P.**), Recall (**R.**), and averaged **F1** scores based on sequence alignment. All metrics are reported in percentages (%).

5 Results and Discussion

5.1 Main Results

Table 1 reports results on the **Synthetic-JZ** and **Real-JZ** benchmarks. We compare JZ-Glyph with frozen API-based VLMs, fine-tuned OCR methods, and variants of JZ-Glyph using different open-source VLM backbones under identical supervision. Unless otherwise specified, JZ-Glyph adopts Qwen3-VL-8B as the default backbone.

Data & framework effectiveness. We first evaluate strong proprietary VLMs with frozen parameters, including GPT-5.1, Qwen-VL (Max / 3-Plus), and Gemini (2.5-Flash / 3-Pro). Despite their strong performance on generic vision–language benchmarks, these models perform poorly on Jianzi recognition. As *Jianzi glyphs visually resemble Chinese characters but follow a distinct and unbounded structural system*, frozen VLMs

Settings	F1↑	CER↓	Acc↑
JZ-Glyph	**84.20**	**19.45**	**63.02**
w/o Style Recomposition	78.77	28.41	52.63
w/o Noise Modeling	78.03	27.10	53.48
w/o Legality-guided Correction	82.68	21.54	60.03
w/o All Components	74.13	31.72	46.13

Table 2: Ablation Study. Performance of JZ-Glyph under different component removal settings, evaluated on Real-JZ. *w/o All* denotes training on raw data directly constructed from metadata, without any augmentation.

Group	Prop.	TAR↓	SAR↓	F1↑	CER↓	Acc↑
All	100.0	18.7	41.3	82.68	21.54	60.03
No anomaly	54.38	0.0	0.0	93.83	7.61	81.76
≥1 anomaly	45.62	53.76	100.0	69.40	38.14	34.11

Table 3: Structural Legality Analysis on Real-JZ. Performance grouped by legality-based anomaly statistics. **Prop.**, **TAR**, and **SAR** denote sample proportion and token-/sequence-level anomaly rates.

may hallucinate familiar characters and produce structurally invalid outputs. On Real-JZ, most frozen models achieve below **21%** accuracy with high CER; even the best model, Gemini-3-Pro, reaches only **27.91%** accuracy and **59.08%** F1. In contrast, when trained with our synthetic data and framework, JZ-Glyph improves accuracy by **+67.95** and **+35.11** on Synthetic-JZ and Real-JZ, respectively, demonstrating that the proposed data and JZ-Glyph framework substantially enhance Jianzi recognition. We further analyze data efficiency in Appendix F, showing that strong performance can already be achieved with only a small fraction of the training data.

Data effectiveness under different paradigms. To isolate the effect of our synthetic data from modeling choices, we reproduce two representative OCR-style paradigms for Jianzi recognition: **CCR-CLIP** (Yu et al., 2023), a CLIP-based retrieval framework aligning glyphs with IDS-like structures, and **YOLO-Radical** (Hayami et al., 2025), which treats Jianzi recognition as fixed-class radical detection. In addition, we include two strong sequence-generation baselines, **TrOCR** (Li et al., 2023) and **PARSeq** (Bautista and Atienza, 2022), which are designed for visual sequence recognition. Implementation details are provided in Appendix B, C and D.

Under the same **Synthetic-JZ** supervision, OCR-style methods perform reasonably well on Synthetic-JZ (**67.48%** and **48.46%** accuracy), confirming that our data provides effective training signals even for conventional approaches. Sequence-generation baselines achieve higher performance (e.g., **82.48%** for TrOCR), reflecting their stronger modeling capacity for structured outputs. However, all baselines degrade substantially on the real-manuscript **Real-JZ** benchmark: OCR-style methods drop to **13.89%** and **5.94%** accuracy with extremely high CER, while sequence-generation models also suffer a large performance gap (e.g., TrOCR **45.58%**).

This gap exposes the fundamental limitations of closed-set, retrieval- or detection-based OCR paradigms and generic sequence-generation models when applied to Jianzi's open-ended and unbounded compositional system, motivating the need for a structure-aware vision-to-sequence formulation for real-manuscript transcription.

Backbone generalization. We further instantiate JZ-Glyph with multiple VLM backbones. Fine-tuning open-source VLMs yields substantial improvements over frozen models on both benchmarks, with a clear backbone scaling trend: stronger models such as InternVL3.5-8B and Qwen3-VL-8B consistently outperform earlier variants, highlighting the importance of model capacity for learning compositional structure.

Overall, although all methods perform better on Synthetic-JZ than on Real-JZ, JZ-Glyph consistently generalizes across different VLM backbones and transfers effectively from synthetic supervision to real-world manuscripts.

5.2 Ablation Study

We conduct an ablation study on Real-JZ to quantify the contribution of each component in **JZ-Glyph** (Table 2).

- Removing *component-wise style recomposition* causes a clear performance drop (**−5.43** F1, **−10.39** Acc, **+8.96** CER), highlighting its importance for bridging the appearance gap between synthetic data and real manuscripts.
- Disabling *manuscript-domain noise modeling* also leads to notable degradation (**−6.17** F1, **−9.54** Acc), indicating that realistic capture and degradation effects are essential for generalization to real-world Jianzi Pu sources.
- Removing the *legality-guided correction module* results in a smaller but consistent decline

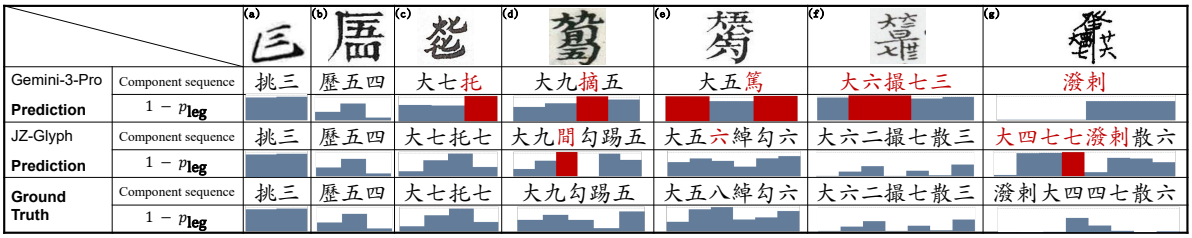

Figure 4: Case study across seven Jianzi samples of increasing structural complexity. Each prediction is shown with its $1 - p_{\text{leg}}(v_t \mid c_t)$ score; red bars mark components flagged as structurally illegal.

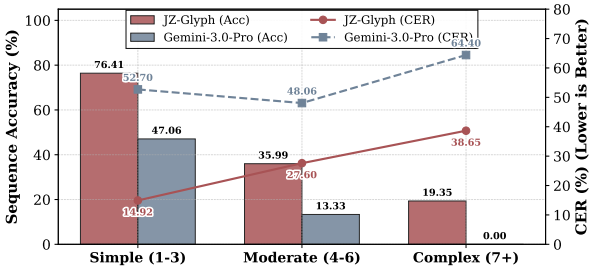

Figure 5: Effect of Structural Complexity. Performance of JZ-Glyph and Gemini-3-Pro across Simple, Moderate, and Complex Jianzi groups.

across metrics, confirming its role in improving structural validity at inference time.

• Finally, removing *all augmentation components* and training on raw synthetic data yields substantial performance loss, demonstrating that the proposed data synthesis pipeline is critical for transferring to real Jianzi manuscripts.

5.3 Legality-based Anomaly Analysis

We analyze structural anomalies on Real-JZ using predictions from the Qwen3-VL-8B backbone before legality-guided correction (Table 3). In Table 3, **Prop.** denotes the sample proportion, **TAR**↓ the fraction of components with legality scores below a fixed threshold, and **SAR**↓ the fraction of sequences containing at least one such component.

Using the legality detector, we find that **18.7%** of components and **41.3%** of sequences exhibit structural anomalies on Real-JZ. These anomalies strongly correlate with recognition failures: anomaly-free sequences achieve **7.61%** CER and **81.76%** accuracy, while anomalous sequences degrade to **38.14%** CER and **34.11%** accuracy.

Overall, this analysis shows that structural legality provides an effective signal for identifying error-prone predictions and supports the use of legality-guided correction to suppress hallucinated components during inference.

5.4 Structural Complexity Analysis

Figure 5 compares performance across *Simple*, *Moderate*, and *Complex* Jianzi groups, defined by the number of semantic components they contain. As structural complexity increases, Gemini-3-Pro exhibits sharp declines in sequence accuracy and substantial increases in CER, revealing clear limitations in handling highly compositional Jianzi. In contrast, JZ-Glyph degrades much more gracefully and maintains significantly stronger performance in the *Complex* group. This widening performance gap indicates that JZ-Glyph more effectively captures component-level compositional regularities, enabling robust generalization as structural complexity increases.

5.5 Case Study and Error Analysis

We present a case study comparing JZ-Glyph with the strongest baseline, Gemini-3-Pro. Figure 4 shows seven examples of increasing structural complexity, together with each model's predicted component sequences and the scores produced by the legality correction module.

Gemini-3-Pro performs well on simple Jianzi, where coarse visual cues suffice. However, as complexity increases, its predictions increasingly exhibit **component confusions** (d,e) and **omissions** (c,f), with failures becoming most severe for highly compositional symbols (g). By contrast, JZ-Glyph consistently produces more accurate and structurally coherent sequences, even for challenging multi-component cases (f). Remaining errors are mainly caused by **early-style variants** (d) and **visually uncommon component combinations** (e,g), both of which introduce local ambiguity.

Finally, the legality-guided correction module goes beyond detection by re-ranking decoding candidates to suppress structurally implausible components, yielding more coherent predictions without modifying the backbone model.

6 Conclusion

We study Jianzi recognition under a zero-shot setting and formulate it as structured visual information extraction over an open and compositional tablature system. By constructing a component-aligned corpus, applying component-wise style recomposition and manuscript-domain noise modeling, and incorporating a legality-guided correction module that re-ranks decoding candidates, JZ-Glyph enables effective vision-to-sequence learning and suppresses structural hallucinations. More broadly, our results reveal the limitations of current VLMs on unbounded compositional symbolic systems. Jianzi represents a broader class of low-resource performance tablatures with similar structural logic, such as Gongche and Shakuhachi notation. We hope this work supports scalable digitization of historical music manuscripts and contributes to the preservation of cultural heritage.

Limitations

While JZ-Glyph demonstrates the feasibility of zero-shot Jianzi recognition and provides a principled pipeline for compositional data construction and modeling, several limitations remain.

First, a residual distribution gap persists between synthetic training data and real-world manuscripts, particularly for rare early-style variants, extreme calligraphic deformation, and severe physical degradation, which can still lead to recognition errors.

Second, the legality-guided module performs correction through candidate re-ranking based on local structural priors, but its corrective capacity is inherently limited. Joint decoding mechanisms may further improve robustness.

Finally, this work focuses on isolated Jianzi recognition. Extending the framework to full-page Jianzi Pu layout understanding, multi-symbol interaction, and complete tablature transcription remains an important direction for future research.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions, which have greatly improved this paper.

We also thank the volunteer annotators for their efforts in constructing the Real-JZ dataset, and Ms. Zhang, a professional Guqin performer, for her adjudication during annotation and her valuable guidance on Jianzi Pu interpretation.

We are grateful to the Guanglingsan community platform for providing the composition data that serves as the primary source for our dataset construction. We would like to express our special thanks to Mr. Wenming Zhang, the key developer of the Guangling Shenqi tool, whose work played a crucial role in enabling the digitization and processing of Jianzi Pu.

This work is supported by the National Natural Science Foundation of China (62276057).

References

Xiang Ao, Xiao-Hui Li, Xu-Yao Zhang, and Cheng-Lin Liu. 2025. Bayesian classifier calibration based on synthesized samples for zero-shot chinese character recognition. *Pattern Recognition*, page 112251.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xiaoyi Bao, Zhongqing Wang, Jinghang Gu, and Chu-Ren Huang. 2025a. Calligraphicocr for chinese calligraphy recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4865–4877.

Xiaoyi Bao, Zhongqing Wang, Jinghang Gu, and Chu-Ren Huang. 2025b. Revisiting classical chinese event extraction with ancient literature information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8440–8451.

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer.

Ruwei Dai, Chenglin Liu, and Baihua Xiao. 2007. Chinese character recognition: history, status and prospects. *Frontiers of Computer Science in China*, 1(2):126–136.

Haisu Guan, Huanxin Yang, Xinyu Wang, Shengwei Han, Yongge Liu, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024. Deciphering oracle bone language with diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15554–15567.

Meguru Hayami, Shun Kuremoto, Mamiko Koshiba, Takashi Kuremoto, and Shingo Mabu. 2025. Recognition of radicals of guqin music notation by yolos. In *International Conference on Multimedia Information Technology and Applications*, pages 118–124. Springer.

Yu-Fen Huang, Jeng-I Liang, I-Chieh Wei, Li Su, and 1 others. 2020. Joint analysis of mode and playing

technique in guqin performance with machine learning. In *ISMIR*, pages 85–92.

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

Mei-Yen Lee. 2023. Concept of nature in the musical aesthetics of the chinese guqin. *Journal of Comparative Literature and Aesthetics*, 46(1):161–171.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Enzhi Ni, Minjun Jiang, and Changle Zhou. 2010. Decomposition of reduced character of chinese guqin notation. In *2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering*, pages 384–389. IEEE.

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

Sina Semnani, Han Zhang, Xinyan He, Merve Tekgürler, and Monica Lam. 2025. Churro: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34765–34812.

Chen Shi. 2016. Guqin notation and music style recognition. *Computer Science*.

Kaiwen Shi, Kan Liu, and Xizi Zhang. 2024. Guqin jianzi notation transcription based on language model. *Communications in Computer and Information Science*, 2007:66–79.

Tom Simon, William Mocaer, Pierrick Tranouez, Clément Chatelain, and Thierry Paquet. 2025. Classifying the unknown: In-context learning for open-vocabulary text and symbol recognition. In *International Conference on Document Analysis and Recognition*, pages 224–243. Springer.

Mingni Tang, Jiajia Li, Lu Yang, Zhiqiang Zhang, Jinhao Tian, Zuchao Li, Lefei Zhang, and Ping Wang. 2025. Nota: Multimodal music notation understanding for visual large language model. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7160–7173.

Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. 2025. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19921–19930.

Peng Wang, Shuai Bai, Sinan Tan, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiao-Feng Wang, Zhi-Huang He, Kai Wang, Yi-Fan Wang, Le Zou, and Zhi-Ze Wu. 2023. A survey of text detection and recognition algorithms based on deep learning technology. *Neurocomputing*, 556:126702.

Bing Wei and Youdi Wang. 2023. Advanced digitization for ancient chinese guqin scores based on mask r-cnn algorithm. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 370–375. IEEE.

Zhenlin Xu, Yi Zhu, Siqi Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. 2024. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1827–1836.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bowen Yang, Shun Kuremoto, Mamiko Koshiba, Shingo Mabu, Hiroo Hieda, and Takashi Kuremoto. 2023. A guqin notation recognition system using machine learning methods.

Zhenhua Yang, Dezhi Peng, Yuxin Kong, Yuyi Zhang, Cong Yao, and Lianwen Jin. 2024. Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6603–6611.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, and 1 others. 2025b. Cc-ocr: A comprehensive and challenging ocr

benchmark for evaluating large multimodal models in literacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21744–21754.

Haiyang Yu, Xiaocong Wang, Bin Li, and Xiangyang Xue. 2023. Chinese text recognition with a pre-trained clip-like model through image-ids aligning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11943–11952.

Minjing Yu, Meng Zhang, Chun Yu, Xiaoguang Ma, Xing-Dong Yang, and Jiawan Zhang. 2021. We can do more to save guqin: Design and evaluate interactive systems to make guqin more accessible to the general public. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Can Zhang, Ziheng Wu, Zhenghao Chen, Yufei Zhan, Yifan Li, Zhao Zhang, Xian Wang, Minghui Qiu, and 1 others. 2025a. Seeing is believing? mitigating ocr hallucinations in multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yuyi Zhang, Yongxin Shi, Peirong Zhang, Yixin Zhao, Zhenhua Yang, and Lianwen Jin. 2025b. Mega-han97k: A large-scale dataset for mega-category chinese character recognition with over 97k categories. *Pattern Recognition*, page 111757.

Yuyi Zhang, Yuanzhi Zhu, Dezhi Peng, Peirong Zhang, Zhenhua Yang, Zhibo Yang, Cong Yao, and Lianwen Jin. 2025c. Hiercode: A lightweight hierarchical codebook for zero-shot chinese text recognition. *Pattern Recognition*, 158:110963.

Shubin Zhou, Fanshuang Meng, and Yanmei Huang. 2025. Semantic modeling of ancient chinese guqin books. *Knowledge Organization*, 52(3):38541.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Dataset	Split	#Jianzi	#Unique Seq.	Simple (%)	Moderate (%)	Complex (%)
Synthetic-JZ	Train	24,198	20,170	29.03	55.92	15.05
	Dev	3,025	2,407	32.07	52.69	15.24
	Test	3,025	2,556	27.77	55.83	16.40
Real-JZ	Test	2,339	602	65.41	33.26	1.33

Table 4: **Dataset statistics for the Synthetic-JZ and Real-JZ benchmarks.** "#Jianzi" counts the total number of segmented symbols, and "#Unique Seq." refers to the number of distinct canonical notation sequences. Complexity levels (*Simple/Moderate/Complex*) are defined by the number of semantic components per Jianzi.

A Dataset Statistics and Manual Annotation Details

Dataset Statistics. Table 4 summarizes the statistics of the *Synthetic-JZ* and *Real-JZ* benchmarks used throughout our experiments. We adopt a zero-shot split where no canonical Jianzi sequence appears in more than one partition. For structural analysis, each symbol is categorized as *Simple*, *Moderate*, or *Complex* according to the number of semantic components (Figure 2).

Manual Annotation Details. Real-JZ is a manually annotated benchmark curated from carefully selected historical *Jianzi Pu* facsimiles (see Appendix E), aiming to capture real manuscript variability such as handwriting styles, ink bleed, and page degradation. Each sample corresponds to a single segmented Jianzi glyph image and is annotated with its canonical component sequence $y = (v_1, \ldots, v_T)$, where every token v_t is drawn from the unified component vocabulary $\mathcal{V}$ used throughout our framework. This design enables consistent sequence-level and token-level evaluation under the same output space. Annotation is conducted using a web-based labeling system (**interface shown in Figure 9**), which supports glyph boxing and constrained token input: annotators entered component tokens via auto-completion over $\mathcal{V}$, together with built-in validity checks to prevent out-of-vocabulary tokens and malformed sequences. The system additionally logs annotator IDs and full revision histories for traceability.

In our annotation setup, three volunteer annotators with Jianzi reading experience (Guqin practitioners) performed the main labeling, and one professional Guqin performer served as the adjudicator. The annotation took approximately one week, totaling about 20 person-hours (0.5 minute per glyph on average, including review). We follow a calibration–annotation–adjudication workflow. The volunteer annotators first label a small calibration subset to align conventions for component decomposition, ordering, and normalization of common calligraphic variants to canonical components. During formal annotation, each glyph is labeled by a single volunteer annotator; ambiguous cases are explicitly flagged rather than force-labeled. For quality control, we automatically run consistency checks over all annotations (e.g., sequence-format validation and mismatch detection against canonical component inventories). All flagged cases and any detected inconsistencies are then reviewed and adjudicated by the professional performer to produce the final labels.

B Reproduction Details of CCR-CLIP Baseline

We reproduce the first-stage pretraining paradigm of CCR-CLIP (Yu et al., 2023) as a representative OCR baseline. The original method is motivated by the observation that Chinese characters admit closed and standardized Ideographic Description Sequences (IDS) defined by Unicode, enabling recognition via image–IDS matching rather than closed-set classification. This formulation, however, implicitly assumes a fixed and enumerable component vocabulary.

In our reproduction, only the stage-1 dual-encoder model is trained from scratch. Jianzi images and their corresponding component sequences are mapped into a shared 2048-dimensional embedding space using CLIP-style contrastive learning. Component sequences are decomposed at the sub-symbol level, padded to a maximum length of 64, and encoded by a Transformer text encoder (6 layers, width 512, 8 heads), followed by a linear projection. Images are resized to 128×128 and encoded by a ResNet-50 backbone with a linear projection to the same embedding space. Training is performed using AdamW with a learning rate of $1\mathrm{e}{-4}$, batch size 16, and a temperature-scaled similarity objective.

Since the component space of Jianzi is open and unbounded, closed-set retrieval over a predefined IDS lexicon is infeasible. We therefore retrieve the nearest sequence from the full **Synthetic-JZ** dataset as an approximate reference, which does not constitute true zero-shot learning but provides a meaningful OCR baseline for comparison.

C Reproduction Details of YOLO-Radical Baseline

We reproduce **YOLO-Radical** (Hayami et al., 2025) as a representative detection-based OCR baseline for Jianzi recognition. The original method treats Guqin Jianzi recognition as fixed-class radical/component detection using a YOLO-style detector, and then aggregates detected radicals to infer the target symbol. This formulation is inherently closed-set and provides limited support for Jianzi's open-ended composition, but it offers a meaningful comparison point for conventional OCR paradigms.

We construct a detection dataset from our Synthetic-JZ metadata (§3.3), where each glyph is associated with a set of component instances and their bounding boxes. Specifically, we collect all component types appearing in the metadata field as the detection label space. For each glyph image, we convert every component instance with pixel-level box annotations into the standard YOLO format $\langle c, x_{center}, y_{center}, w, h \rangle$ normalized by the image width and height.

In the training phase, we implement YOLO-Radical using the **YOLO11** detector. We initialize from a pretrained checkpoint and train for 100 epochs with image size 128 and batch size 16. We use standard YOLO training settings provided by the Ultralytics framework.

In the Inference phase, given a test glyph image, the detector outputs a set of component boxes with class predictions. Since the YOLO-Radical paradigm does not model the ordered component sequence, we follow a simple **bag-of-components** decoding scheme: we count predicted component classes to obtain a multiset, and match it to candidate glyphs by comparing multiset with each glyph's ground-truth component multiset derived from metadata. We choose the best-matching glyph ID by maximizing a multiset F1 score, and output the corresponding canonical component sequence text associated with that glyph ID. This yields a deterministic closed-set decoding from detections to a sequence string.

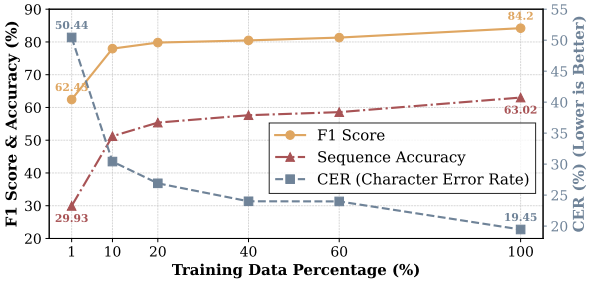

Figure 6: **Data Efficiency.** Performance of JZ-Glyph under different fractions of training data.

D Reproduction Details of TrOCR and PARSeq

For TrOCR, we use trocr-base-stage1 with a learning rate of 2e-5 for 15 epochs, applying early stopping. For PARSeq, since the released pretrained weights are built for an English-oriented charset, we construct the charset from the full training set and train with a learning rate of 7e-4, a batch size of 64, for 30 epochs, also with early stopping.

E Facsimile Examples from Jianzi Pu

To illustrate the diversity and real-world complexity of `Guqin Jianzi Pu` manuscripts, we present facsimile pages from two representative sources (Figure 7, 8). These historical materials also constitute the primary sources from which our **Real-JZ** evaluation set is curated. The examples exhibit substantial stylistic variation, handwritten irregularities, printing artifacts, and paper degradation—factors that make Jianzi recognition fundamentally different from conventional OCR and highlight the distribution gap between synthetic training data and real manuscripts. For clarity, all Jianzi symbols in the images are marked with red bounding boxes.

F Data Efficiency Analysis

Figure 6 shows that using only **10%** of the training data (approximately 2.4k samples), JZ-Glyph already achieves strong recognition performance on the Real-JZ benchmark. Increasing the data size yields consistent improvements in F1, CER, and sequence accuracy, but with clear *diminishing returns*: 20–40% of the data recovers most of the full-data performance. This trend indicates that JZ-Glyph learns transferable component-level and structural cues early in training, rather

than memorizing specific glyphs. Component-wise style recomposition increases visual diversity, while manuscript-domain noise modeling reduces the domain gap, together enabling robust generalization from limited supervision to real manuscripts.

G System Prompt

System Prompt

```
你是古琴減字譜數位化專家，請完成減字圖片到譜式組合序列的映射。
# 輸出格式
{
  "character_analysis": [
    {
      "original_character": "視覺描述",
      "components": ["部", "件"],
      "ids": "⿰AB",
      "interpretation": "語義解讀"
    }
  ],
  "sequence": ["泛起", "大八注勾三", "潑剌名七六六散五", ...]
}
# 【部件庫】圖片中可能包含的部件 ** 右手指法部件 **：木(抹)、乚(挑)、勹(勾)、易(剔)、丁(打)、啇(摘)、刀(劈)、乇(托)、早(撮)、厂(歷) ...
** 左手指法部件 **：大、(亻) 食、中、夕(名)、足(跪)、卜(綽)、氵(注) ...
** 弦序與徽位部件 **：一、二、三、四、五、六、七、八、九、十
** 音色部件 **：艹(散)、宀(按)、乏(泛)、己(起)、止...
** 譜法部件 **：從、豆(頭)、乍(作)、再、冬(終)、曲...
** 表情部件 **：尤(就)、勹(急)、虍(虛)、爰(緩) ...
# 分析流程 1. 對照部件庫辨識可見部件 2. 描述原始字形視覺特徵 3. 分解部件，準確辨識數字 4. 構建 IDS 表達式 5. 將結構翻譯為譜式術語
# 關鍵原則 1. ** 數字優先 **：確保數字辨識準確 2. ** 部件對照 **：參考部件庫辨識漢字結構 3. ** 先形後義 **：先分析結構，再解讀指法 4. ** 一一對應 **：character_analysis 與 sequence 順序一致 5. ** 數字一致 **：components 中的數字必須與 sequence 中的數字完全一致 6. ** 單一數字 **：圖片中只有一個數字時，components 中只輸出該數字
# 範例
{
  "character_analysis": [
    {
      "original_character": "上下結構，艹頭+勹形+單橫線",
      "components": ["艹", "勹", "一"],
      "ids": "⿱艹⿰勹一",
      "interpretation": "散音勾一弦"
    },
  ],
  "sequence": ["散勾一"]
}

{
  "character_analysis": [
    {
    "original_character": "左右結構，夕部+九字+乚形+七字",
      "components": ["夕", "九", "乚", "七"],
      "ids": "⿰夕⿰九⿰乚七",
      "interpretation": "名指九徽挑七弦"
    }
  ],
  "sequence": ["名九挑七"]
}
開始辨識：
```

System Prompt (English Translation)

```
You are an expert in Guqin Jianzi pu (reduced notation) digitization. Please map the Jianzi image to a notation combination sequence.
# Output Format
{
  "character_analysis": [
    {
      "original_character": "Visual Description",
      "components": ["Component", "Component"],
      "ids": "⿰AB",
      "interpretation": "Semantic Interpretation"
    }
  ],
  "sequence": ["泛起", "大八注勾三", "潑剌名七六六散五"]
}
# [Component Library] Visual parts likely present **Right-Hand Fingering**: 木(抹)、乚(挑)、勹(勾)、易(剔)、丁(打)、啇(摘)、刀(劈)、乇(托)、早(撮)、厂(歷) ...
**Left-Hand Fingering**: 大、(亻) 食、中、夕(名)、足(跪)、卜(綽)、氵(注) ...
**String Order & Hui Position**: 一, 二, 三, 四, 五, 六, 七, 八, 九, 十 (1-10)
**Timbre**: 艹(散)、宀(按)、乏(泛)、己(起)、止...
**Notation Method**: 从、豆(頭)、乍(作)、再、冬(終)、曲...
**Expression**: 尤(就)、勹(急)、虍(虛)、爰(緩) ...
# Analysis Process 1. **Identify Components**: Match visible parts against the Component Library. 2. **Describe Visuals**: Describe the visual characteristics of the original character. 3. **Decompose**: Break down parts and accurately identify numbers. 4. **Construct IDS**: Build the Ideographic Description Sequence (IDS). 5. **Translate**: Convert the structure into musical notation terminology.
# Key Principles 1. **Numbers First**: Ensure absolute accuracy in number identification. 2. **Component Reference**: Strictly refer to the library for character structure identification. 3. **Form Before Meaning**: Analyze the structure first, then interpret the fingering semantics. 4. **One-to-One Correspondence**: The order of ‘character_analysis‘ must match ‘sequence‘. 5. **Number Consistency**: Numbers in ‘components‘ must exactly match numbers in ‘sequence‘. 6. **Single Number**: If the image contains only one number, output only that number in ‘components‘.
# Examples
{
  "character_analysis": [
    {
    "original_character": "Top-bottom structure, 艹 header + 勹 shape + single horizontal line",
      "components": ["艹", "勹", "一"],
      "ids": "⿱艹⿰勹一",
      "interpretation": "Open string (San), Hook (Gou) on String One"
    }
  ],
  "sequence": ["散勾一"]
}

{
  "character_analysis": [
    {
    "original_character": "Left-right structure, 夕 part + Nine + 乚 shape + Seven",
      "components": ["夕", "九", "乚", "七"],
      "ids": "⿰夕⿰九⿰乚七",
    "interpretation": "Ring finger at Hui 9, Tiao on String 7"
    }
  ],
  "sequence": ["名九挑七"]
}
Start Recognition:
```

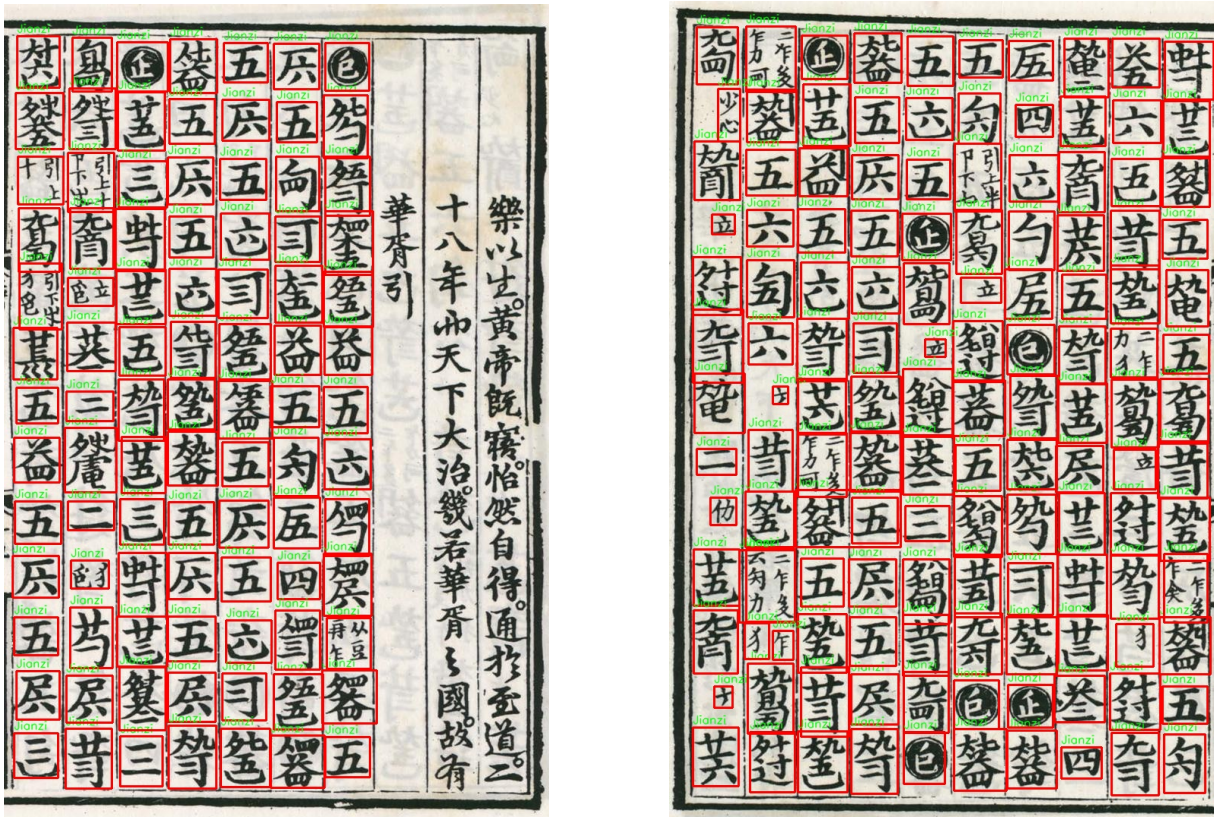

Figure 7: Facsimile sheets of *"Huaxu Yin" from the Shenqi Mipu* (《神奇祕譜·華胥引》), first published in 1425 and compiled by Zhu Quan. All Jianzi (減字) in the images are highlighted with red bounding boxes.

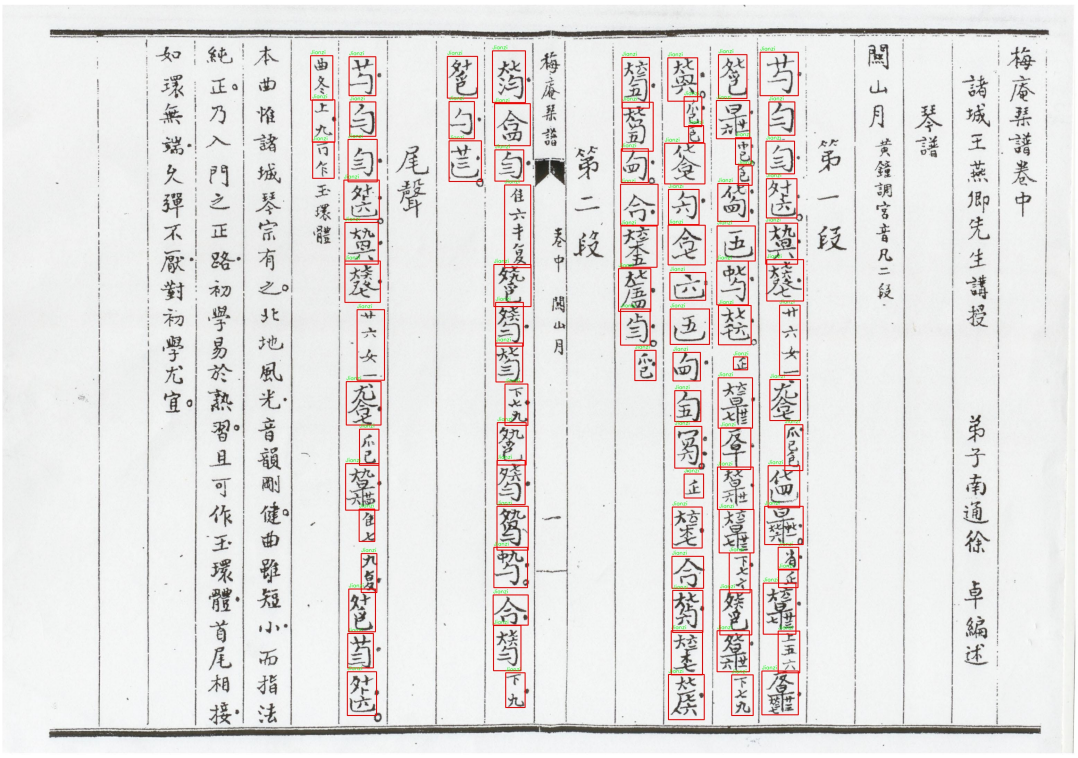

Figure 8: A facsimile sheet of *"Guanshan Yue" from the Mei-an Qinpu* (《梅庵琴譜·關山月》), first published in 1931 and compiled by Xu Lisun and Shao Sen. All Jianzi (減字) in the image are highlighted with red bounding boxes.

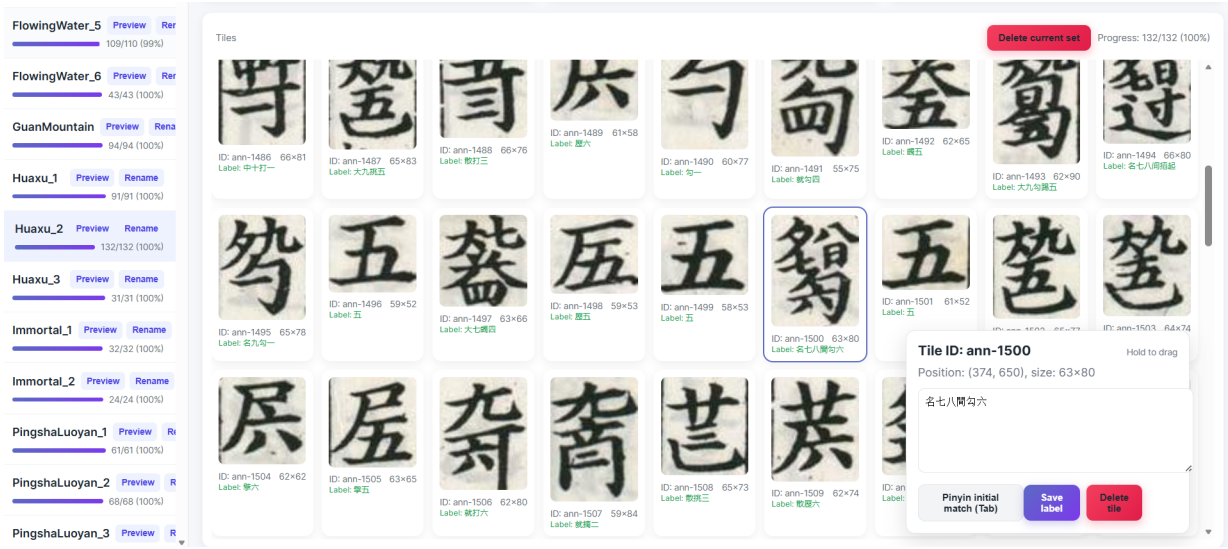

Figure 9: **Web-based annotation interface for Real-JZ.** The system supports efficient glyph-level labeling on manuscript crops: the left panel lists manuscript sets and progress, the center shows the cropped Jianzi tiles in a grid with IDs and current labels, and the right panel provides an editable form for the selected tile (bounding box metadata, component-sequence label entry, and shortcut actions) with revision logging for traceability.