DOMAIN-BOOTSTRAPPED OCR FOR LOW-RESOURCE TRADITIONAL CHINESE PAINTING INSCRIPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional Chinese Painting Inscriptions (TCPIs) carry critical information for authorship attribution, artwork dating, provenance tracing, and cultural interpretation, making automatic recognition of such inscriptions an important task for OCR in cultural heritage research. However, TCPIs pose challenges beyond standard document or scene-text OCR due to severe domain shifts, including free-form brush calligraphy, complex background interference, faded ink, and other degradations. These challenges are amplified in low-resource settings, where high-quality annotated TCPI data are scarce and expensive to obtain. To address these issues, we propose a domain-bootstrapped OCR pipeline tailored for low-resource TCPIs. Original paintings are divided into small and large groups and cropped into TCPI patches. The smaller group is used to construct a high-quality seed dataset through multi-model candidate generation and expert verification. A general-purpose OCR model is fine-tuned on this seed dataset to produce an initial TCPI-adapted model, which then generates pseudo-labels for the larger unlabeled group. The same OCR model is subsequently retrained via bootstrapped training on these pseudo-labeled samples using optimized strategies to further enhance recognition performance. Experimental results show that the TCPI-bootstrapped model achieves a clear performance advantage ($F1 = 0.851$) over existing state-of-the-art methods. Ablation studies further demonstrate that recognition performance consistently improves as the pseudo-labeled dataset expands, highlighting the effectiveness and scalability of the proposed domain-bootstrapping pipeline. The model and benchmark will be released upon acceptance.

1 INTRODUCTION

Traditional Chinese Painting Inscriptions (TCPIs) are textual elements in Traditional Chinese Paintings (TCPs), including appreciation poems, artist signatures, artwork dates, colophons, and other annotations. A representative example is provided in Appendix A. They provide indispensable evidence for dating artworks, tracing provenance, and conducting iconographic analysis. With the rapid growth of digital museum collections and online image repositories, automatic recognition of TCPIs has become a critical task in intelligent cultural heritage processing.

TCPIs display extreme visual and linguistic variability. Calligraphy can take multiple styles, including regular kaishu, running xingshu, cursive caoshu, clerical lishu, and seal zhuanshu, which often coexist within a single work. Strokes are executed freely and vary in ink density, dryness, pressure, and continuity. Text layouts are highly flexible and frequently interwoven with visual elements such as rocks, brush textures, moss dots, and color washes, making text boundaries difficult to delineate. Ink fading, diffusion, and the aging of paper or silk further reduce stroke clarity and increase background noise. These factors make TCPIs difficult to annotate, requiring specialized expertise and limiting the availability of high-quality labeled data. Collectively, these characteristics make TCPIs particularly challenging for automatic text recognition.

Existing OCR methods are primarily designed for printed documents, natural scene text, or single-script handwriting, relying on regular layouts, clear strokes, and stable glyphs (Cui et al., 2026a; Wu et al., 2026; Duan et al., 2026; Team et al., 2025). Approaches developed for historical manuscripts or inscriptions generally assume relatively orderly layouts and controlled noise (Semnani et al., 2025; Meoded, 2025; Wraight et al., 2025), which restricts their applicability to TCPIs. Recent advances in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

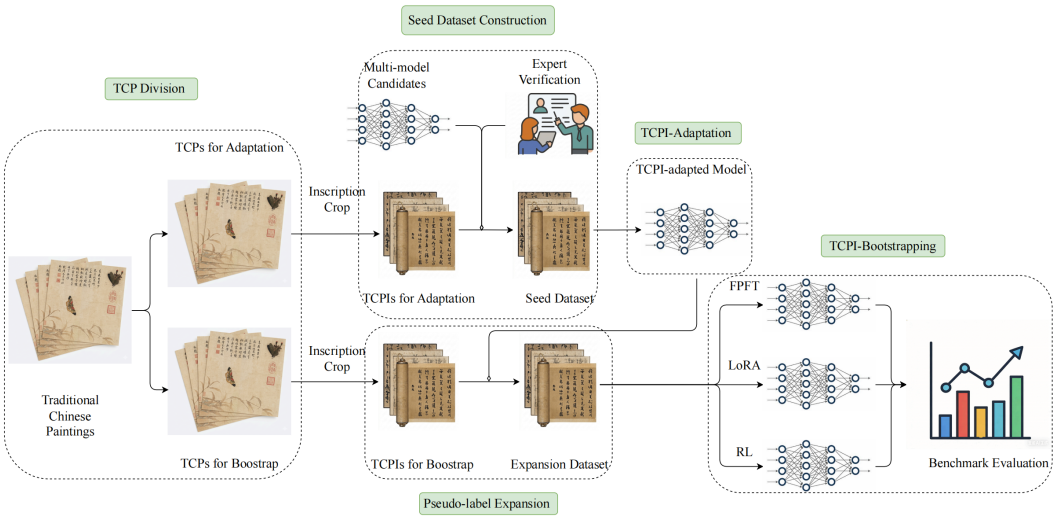

Figure 1: Overview of our domain-bootstrapped OCR pipeline for TCPIs.

vision-language models have shifted OCR from conventional detect-recognize-postprocess pipelines to unified frameworks that jointly model text and visual context (Poznanski et al., 2025; Li et al., 2025a; Yu et al., 2025). These models can handle complex layouts, multi-script documents, and text interwoven with images, but they typically require large amounts of training data, limiting their effectiveness in low-resource scenarios such as TCPIs (Gagnier et al., 2026; Lin et al., 2026). As a result, automatic recognition of TCPIs remains an open and challenging problem.

To address these challenges, we propose a domain-bootstrapped OCR pipeline for TCPIs, guided by the principle of leveraging a small set of high-quality expert annotations and progressively expanding through bootstrapping. The overall pipeline architecture is shown in Figure 1.

The pipeline divides paintings before cropping, constructs a human-verified seed set from multiple OCR candidates, adapts a general-purpose OCR model, filters its pseudo-labels on a larger unlabeled collection, and retrains on the expanded supervision. Section 3 defines these five modules in detail.

Experiments show that our pipeline significantly outperforms other SOTA OCR models and LLMs in character-wise metrics. Ablation studies confirm the effectiveness of seed dataset construction, pseudo-label expansion, and bootstrapped training, highlighting the scalability of the approach.

Our main contributions are as follows:

1. We introduce a systematic domain-bootstrapping paradigm into the extremely low-resource and high-noise domain of TCPIs. By orchestrating multi-model seed construction, pseudo-label expansion, and bootstrapped retraining, we effectively bridge the gap between general-purpose OCR architectures and the unique visual-linguistic complexities of TCPIs.
2. We construct and release the first standard evaluation benchmark for TCPIs. To overcome the prohibitive annotation costs and extreme domain complexities, we formalize a collaborative machine-expert annotation protocol that efficiently mitigates domain shifts and guarantees dataset quality.
3. We release a robust TCPI-bootstrapped model. Extensive experiments and ablation studies demonstrate its superior performance over general-purpose OCR systems and multi-modal LLMs, providing a scalable adaptation pathway for heritage text recognition.

2 Related Work

OCR and document understanding. Modern OCR increasingly uses scaled vision-language models and data-centric training to unify recognition with document structure (Cui et al., 2026a; Wu et al., 2026; Duan et al., 2026; Team et al., 2025; Poznanski et al., 2025; Wei et al., 2026;

Cui et al., 2026b; Tang, 2026). Structured multimodal systems extend this direction to layout and grounded document reasoning (Li et al., 2025a; Dong et al., 2026a; Feng et al., 2026; Yu et al., 2025; Hu et al., 2024; Li et al., 2025b; Curini et al., 2026), supported by increasingly fine-grained benchmarks (Zhu et al., 2024; Zhang et al., 2025; Maeda & Okazaki, 2026). In artistic or degraded imagery, however, unified models can hallucinate structure or substitute semantic guesses for weak visual evidence. Diffusion decoding, geometric verification, model fusion, and consistency learning offer partial safeguards (Dong et al., 2026b; Gong et al., 2026; Tekin et al., 2026; Luo et al., 2025), while comparative studies still find value in explicit OCR grounding or decoupled pipelines (Hasin et al., 2026; Nguyen et al., 2025; Benkirane et al., 2026; Beyene & Dancy, 2026). TCPIs intensify these problems because text strokes are interwoven with non-text brushwork and irregular layouts.

Low-resource adaptation and historical HTR. Historical recognition addresses sparse supervision, costly annotation, and degradation through specialized transcription and modular analysis (Semnani et al., 2025; Meoded, 2025; Wraight et al., 2025; Neji et al., 2026; Le & Kitamoto, 2025; Bassanini et al., 2026). Generic VLM failures on ancient, minority-language, and historically authentic text demonstrate the need for target-specific alignment (Gagnier et al., 2026; Angleraud et al., 2026; Kong et al., 2026; Liu et al., 2026; Vesalainen et al., 2026). Existing remedies include few-shot conditioning, efficient domain adaptation, image enhancement, targeted fine-tuning, and human-in-the-loop corpus construction (de Senneville et al., 2026; Dev & Zhan, 2026; Chendage & Mente, 2026; Tran et al., 2026; Lin et al., 2026; Kermorvant et al., 2024; Jayatilleke et al., 2026; Lasandi & Jayatilleke, 2026). Calligraphy datasets further highlight style and layout diversity (Zhao et al., 2025). Our approach combines these principles: expert arbitration establishes a low-noise seed, and filtered pseudo-labels from real paintings expand coverage of the target distribution.

3 METHODOLOGY

TCPIs carry rich semantic, historical, and stylistic information, but recognizing them from images remains highly challenging. Unlike standard OCR scenarios, TCPI recognition involves complex calligraphic expressions, irregular layouts, and frequent visual interference such as brush strokes, seals, and degradation artifacts.

These characteristics introduce three fundamental challenges. First, supervision is extremely sparse, as accurate annotation requires expert knowledge in calligraphy and classical Chinese, making large-scale labeling prohibitively expensive. Second, there exists a substantial domain shift between TCPI and general OCR data, due to diverse writing styles, non-standard structures, and artistic variations. Third, a non-negligible portion of inscriptions is partially or entirely degraded, making them excessively noisy for standard learning objectives and requiring rigorous data quality control.

Problem Definition. Given an inscription crop image x extracted from a TCP, the TCPI recognition task is formulated as predicting the most probable character sequence:

$$y^* = \arg \max_{(c_1, c_2, \ldots, c_n)} P(c_1, c_2, \ldots, c_n \mid x).$$

To address the limited supervision, we further consider a semi-supervised learning setting. Let $\mathcal{D}_{\text{seed}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a small set of high-quality annotated samples, and $\mathcal{D}_{\text{unlabeled}} = \{x_j\}_{j=1}^{M}$ denote a large collection of unlabeled inscription images. The goal is to learn a recognition model f_θ that can effectively transfer knowledge from $\mathcal{D}_{\text{seed}}$ to $\mathcal{D}_{\text{unlabeled}}$, enabling automatic annotation at scale.

To achieve this, we optimize the following objectives.

(1) Supervised learning on annotated samples. For labeled data $(x, y) \in \mathcal{D}_{\text{seed}}$, we maximize the conditional likelihood:

$$\mathcal{L}_{\text{sup}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{seed}}} \left[\log P_\theta(y \mid x)\right].$$

(2) Learning from unlabeled data via pseudo-labeling. For unlabeled samples $x \in \mathcal{D}_{\text{unlabeled}}$, the adapted model generates candidate transcriptions $\hat{y} = f_\theta(x)$. To control error propagation, we retain only high-confidence predictions based on a specialized generating confidence score $s_{\text{conf}}(\cdot)$:

$$\mathcal{D}_{\text{pseudo}} = \{(x, \hat{y}) \mid s_{\text{conf}}(x, \hat{y}) \geq \tau\}.$$

The model is then further bootstrapped by maximizing:

$$\mathcal{L}_{\text{pseudo}}(\theta) = \mathbb{E}_{(x,\hat{y})\sim\mathcal{D}_{\text{pseudo}}} \left[\log P_\theta(\hat{y} \mid x)\right].$$

The overall optimization progresses from supervised learning on human-verified seed data to bootstrapping on the large-scale expanded corpus.

To address the TCPI recognition problem under sparse supervision, we propose a five-module pipeline that progressively transforms raw paintings into a large-scale pseudo-labeled dataset. The pipeline consists of TCP division, seed dataset construction, TCPI adaptation, pseudo-label expansion, and TCPI bootstrapping.

Given a collection of paintings $\mathcal{P}$, our goal is to learn a recognition model f_θ by constructing and expanding training data from a small annotated subset to a large pseudo-labeled corpus.

3.1 TCP Division

We first transform the painting-level dataset $\mathcal{P}$ into a set of inscription image patches. Specifically, an inscription detector ϕ is applied to each painting $p \in \mathcal{P}$ to localize text regions:

$$\mathcal{X} = \{x = \phi(p) \mid p \in \mathcal{P}\}.$$

To avoid data leakage and ensure distributional diversity, $\mathcal{P}$ is partitioned into two disjoint subsets:

$$\mathcal{P} = \mathcal{P}_{\text{small}} \cup \mathcal{P}_{\text{large}}, \quad \mathcal{P}_{\text{small}} \cap \mathcal{P}_{\text{large}} = \emptyset.$$

The corresponding inscription sets are denoted as $\mathcal{X}_{\text{seed}}$ and $\mathcal{X}_{\text{unlabeled}}$, respectively.

3.2 Seed Dataset Construction

For each $x \in \mathcal{X}_{\text{seed}}$, we obtain multiple candidate transcriptions from a set of OCR models $\{g_k\}_{k=1}^K$:

$$\mathcal{C}(x) = \{g_k(x)\}_{k=1}^K.$$

Human experts then perform selection or correction over $\mathcal{C}(x)$ to produce a high-quality annotation y. This results in a seed dataset:

$$\mathcal{D}_{\text{seed}} = \{(x_i, y_i)\}_{i=1}^N.$$

Unreadable samples are explicitly filtered, preventing noisy supervision.

3.3 TCPI-Adaptation

We initialize a general-purpose OCR model f_{θ_0} and fine-tune it on $\mathcal{D}_{\text{seed}}$ to obtain a TCPI-adapted model:

$$\theta_1 = \arg\max_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{seed}}} \left[\log P_\theta(y \mid x)\right].$$

The resulting model f_{θ_1} captures domain-specific features such as brush-style strokes, mixed scripts, and irregular layouts.

3.4 Pseudo-Label Expansion

The adapted model f_{θ_1} is applied to unlabeled samples $x \in \mathcal{X}_{\text{unlabeled}}$ to generate pseudo-labels:

$$\hat{y} = f_{\theta_1}(x).$$

We retain only high-confidence predictions based on a scoring function $s_{\theta_1}(x)$:

$$\mathcal{D}_{\text{pseudo}} = \{(x, \hat{y}) \mid s_{\theta_1}(x) \geq \tau\}.$$

This step effectively transfers knowledge from $\mathcal{D}_{\text{seed}}$ to a much larger dataset.

Table 1: Dataset composition across training and evaluation stages.

Dataset Split	Samples	Description
Seed train dataset	848	Train dataset for adaptation
Seed validation dataset	94	Validation dataset for adaptation
Seed dataset	942	Total seed dataset
Expansion dataset	4,943	Pseudo-label expansion dataset

3.5 TCPI-BOOTSTRAPPING

We retrain the model on the pseudo-labeled dataset to obtain a bootstrapped model:

$$\theta_2 = \arg\max_{\theta} \mathbb{E}_{(x,\hat{y})\sim\mathcal{D}_{\text{pseudo}}} \left[\log P_\theta(\hat{y} \mid x)\right].$$

The resulting model f_{θ_2} is better aligned with the TCPI data distribution.

The proposed pipeline can be viewed as a progressive data-centric learning process that expands supervision from a small expert-annotated set to a large pseudo-labeled corpus. It effectively bridges the gap between sparse supervision and complex domain-specific distributions in TCPI recognition.

4 EXPERIMENTS

We report the training setup, dataset composition, benchmark evaluation, and ablation study.

4.1 EXPERIMENTAL SETUP

Approximately 12,000 paintings are partitioned into disjoint subsets of 10,000 and 2,000 before a fine-tuned DINO-YOLO model localizes and crops inscriptions. Chandra-OCR-0.1.0 (Paruchuri, 2025), olmOCR-2, and PaddleOCR-VL generate complementary seed candidates. Experts discard entirely indiscernible crops, select an exact candidate when available, and otherwise post-edit or transcribe from scratch, yielding **942** verified samples (Table 1). We split these into **848** training and **94** validation samples and perform full-parameter fine-tuning (FPFT) of olmOCR-2 to obtain the TCPI-adapted model. It labels **5,394** additional crops; a rule-based score penalizes special symbols and repetitions, and the **4,943** predictions scoring at least 0.95 form the expansion set. We retrain on this set with FPFT, Low-Rank Adaptation (LoRA), and Reinforcement Learning (RL). To independently validate the expansion labels, human experts transcribed 463 distribution-matched samples, covering 9.37% of the 4,943-sample expansion set, and compared these references with the generated pseudo-labels. The full protocol and results are reported in Appendix D.

Evaluation Metrics. We evaluate OCR output using character-level precision (P), recall (R), $F1$, normalized edit similarity (NES), and exact match (EM). Before scoring, predictions and references are converted from traditional to simplified Chinese and normalized to remove formatting artifacts. Character-level P, R, and $F1$ measure partial character recovery, NES captures edit-based sequence similarity, and EM requires an exact transcription match. All results are averaged over the evaluated samples. Formal metric definitions are provided in Appendix C.

4.2 SEED DATASET COMPOSITION

To examine the domain characteristics of the seed dataset, we annotate each sample according to the origin of its transcription. A sample is labeled as *OCR-predict* if the final transcription is selected from OCR-generated candidates, and as *Expert-predict* if the transcription is directly provided or corrected by expert annotators. Statistics show that *OCR-predict* samples account for **28.92%** of the dataset, whereas *Expert-predict* samples constitute **71.08%**.

This distribution provides two observations. First, the *OCR-predict* portion indicates that candidate models can partially reduce annotation cost by generating usable or near-correct transcriptions that require only light human arbitration. Second, the dominance of *Expert-predict* samples suggests that a large fraction of TCPIs remain difficult for existing OCR systems due to the domain's high

variability and visual complexity. This further highlights the necessity of domain adaptation for OCR models operating on TCPI data.

To further analyze the diversity of the dataset, we categorize the seed samples along four dimensions: inscription clarity, calligraphic style, visual interference, and text length. Specifically, clarity scores of 1–3 are grouped as degraded and a score of 4 as legible; calligraphic style includes running script, cursive script, and regular script; visual interference covers clean (no significant interference), texture (painting-texture interference), and seal (seal-stamp interference). Calligraphic style and interference are multi-label categories, so one sample may contribute to multiple columns within either dimension. Based on the OCR-body length after front-matter removal, text is grouped as short (0–50 characters), medium (51–100 characters), or long (more than 100 characters). As shown in Figure 2, the dataset exhibits substantial variability across these dimensions, reflecting the heterogeneous characteristics of TCPIs encountered in real-world paintings.

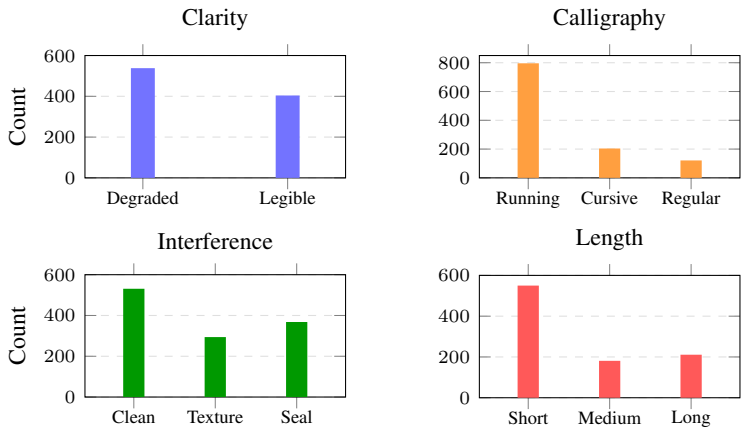

Figure 2: Composition of the seed dataset across four dimensions.

4.3 Benchmark Evaluation

To comprehensively assess recognition performance and alignment robustness, models are evaluated on both the 94-sample zero-shot validation set and the full 942-sample seed dataset. Although the 94-sample validation set provides an unbiased zero-shot evaluation, its small size makes fine-grained metric partitioning highly sensitive to statistical noise. Evaluating the bootstrapped models on the full 942-sample seed set—which was used during the initial adaptation stage—serves as an "In-domain Knowledge Retention Check" (measuring Domain Fidelity). In a self-training pipeline supervised by machine-generated pseudo-labels, models are prone to hallucination or semantic drift. Using the human-verified seed dataset as a stable anchor allows us to confidently verify that the expanding pseudo-labels are reinforcing authentic TCPI distributions rather than deviating from them.

Table 2 presents a performance comparison among the TCPI-adapted model, TCPI-bootstrapped model, the base model olmOCR-2-7B-1025, state-of-the-art OCR systems (Zhang et al., 2026; Li et al., 2026; Wang et al., 2026; Cao et al., 2026), and mainstream multimodal LLMs.

Three key observations emerge from the results. First, the TCPI-bootstrapped model consistently outperforms the base model across both evaluation sets, confirming that the domain-bootstrapping strategy effectively enhances TCPI OCR performance. Second, the TCPI-bootstrapped model improves upon the TCPI-adapted model, indicating that high-quality pseudo-label expansion extends coverage of the target distribution rather than merely repeating training. Third, general-purpose LLMs and open OCR systems remain substantially behind domain-adapted models in this heritage-specific scenario, highlighting the challenges of jointly adapting to complex calligraphic styles, visual interference, and low-resource target distributions.

4.4 Ablation Study

Table 3 summarizes fine-grained *F1* scores across different training scales, augmentation schemes, and training strategies, evaluated on the full 942-sample seed dataset. The baseline is established by

Table 2: Benchmark comparison for TCPI OCR. Validation denotes the 94-sample seed validation set, and Overall denotes the full 942-sample seed dataset.

Model	Validation					Overall				
	P	R	F1	NES	EM	P	R	F1	NES	EM
TCPI-adapted	0.862	0.863	0.861	0.845	**0.223**	/	/	/	/	/
TCPI-bootstrapped	**0.874**	**0.866**	**0.868**	**0.851**	0.202	**0.869**	**0.843**	**0.851**	**0.829**	**0.329**
olmOCR-2-7B-1025	0.828	0.827	0.825	0.803	0.181	0.787	0.770	0.774	0.748	0.203
GLM-OCR	0.739	0.742	0.739	0.725	0.106	0.720	0.723	0.720	0.704	0.126
Chandra OCR 0.1.0	0.750	0.771	0.755	0.714	0.160	0.731	0.750	0.733	0.700	0.170
PaddleOCR-VL-1.6	0.815	0.822	0.816	0.798	0.146	0.797	0.800	0.795	0.777	0.166
FireRed-OCR	0.810	0.802	0.803	0.787	0.160	0.770	0.772	0.767	0.746	0.156
dots.ocr-1.5	0.783	0.773	0.775	0.745	0.149	0.761	0.737	0.745	0.714	0.152
HunyuanOCR-1.5	0.698	0.727	0.699	0.643	0.128	0.740	0.763	0.740	0.693	0.173
MinerU2.5-Pro-2605-1.2B	0.631	0.681	0.632	0.621	0.149	0.608	0.658	0.612	0.592	0.141
Youtu-Parsing	0.630	0.549	0.572	0.512	0.044	0.634	0.560	0.583	0.525	0.088
DeepSeek-OCR-2	0.313	0.225	0.226	0.168	0.011	0.278	0.212	0.211	0.164	0.033
GPT-5.4	0.524	0.564	0.538	0.506	0.064	0.540	0.568	0.548	0.516	0.074
Gemini 3.1 Pro	0.723	0.550	0.586	0.530	0.213	0.703	0.535	0.567	0.512	0.231
Claude Opus 4.6	0.635	0.672	0.641	0.610	0.085	0.621	0.654	0.626	0.589	0.151

Table 3: Fine-grained evaluation on the seed dataset (*F1* score).

Settings	Overall	Clarity		Calligraphy			Interference			Length		
		Degraded	Legible	Running	Cursive	Regular	Clean	Texture	Seal	Short	Medium	Long
Baseline	0.847	0.868	0.819	0.848	0.751	0.929	0.868	0.826	0.817	0.862	0.894	0.766
1/2 Samples	0.835	0.848	0.819	0.836	0.713	0.931	0.853	0.815	0.809	0.852	0.867	0.764
1/5 Samples	0.817	0.832	0.798	0.815	0.699	0.933	0.842	0.794	0.781	0.830	0.847	0.759
Only Rotate	0.819	0.841	0.791	0.818	0.705	0.922	0.844	0.803	0.783	0.835	0.856	0.746
W/o Rotate	**0.851**	0.869	**0.827**	0.850	0.751	**0.936**	**0.869**	0.833	0.822	**0.868**	**0.901**	0.761
LoRA	0.850	**0.871**	0.823	**0.852**	**0.753**	**0.936**	0.865	**0.839**	**0.826**	0.867	0.895	0.768
RL	0.807	0.819	0.791	0.804	0.684	0.933	0.831	0.784	0.772	0.806	0.847	**0.774**

performing FPFT on the complete expanded pseudo-labeled dataset with a comprehensive augmentation suite (Geometric, Degradation, Appearance, and Blur). Throughout these experiments, FPFT and LoRA models are selected based on recall, while RL models are optimized and selected using a verified reward combining soft-overlong, recall, and format scores. The complete reward is defined in Appendix B. Each column corresponds to a major TCPI subgroup along four dimensions: **Clarity**, **Calligraphy**, **Interference**, and **Length**.

Key observations are as follows:

- **Effect of training scale.** Expanding the pseudo-labeled training set consistently improves performance: the overall *F1* score rises from 0.817 (1/5 Samples) to 0.835 (1/2 Samples) and 0.847 (Baseline). This monotonic improvement is further corroborated by the validation trajectories in **Figure 3 (Left)**, where larger sample scales yield steeper learning curves and higher recall ceilings. The result indicates that the model can progressively internalize domain-specific knowledge as high-quality pseudo-labels expand.
- **Impact of augmentation strategies.** Geometric augmentations must be tailored to the domain. Removing rotation augmentation (*W/o Rotate*) raises the overall *F1* score to **0.851**, the highest in the table, whereas using only rotation augmentation yields 0.819. **Figure 3 (Right)** provides additional evidence, showing that the *W/o Rotate* configuration maintains a consistent lead in matching recall from the early stages of training. This suggests that random rotation introduces harmful noise by disrupting the inherent vertical/horizontal layout priors and directional dynamics of TCPI calligraphy.
- **FPFT vs. LoRA vs. RL training strategies.** Parameter-efficient fine-tuning via LoRA achieves a highly competitive overall score of 0.850 and leads on the challenging **Texture** (0.839) and **Seal** (0.826) subsets. RL underperforms in most categories (0.807 overall), but

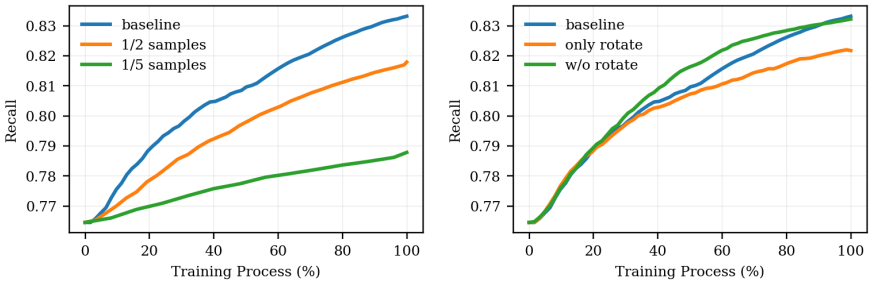

Figure 3: Character-level matching recall curves on the seed validation dataset. Left: Comparison across different training sample scales (1/5, 1/2, and full Baseline). Right: Comparison across different augmentation schemes (Baseline, only rotate, and w/o rotate).

achieves the highest **Long**-text score (0.774), indicating a specialized advantage in extended transcriptions despite weaker general recognition.

- **Advantage of the Bootstrapped Pipeline.** The superior performance of the "W/o Rotate" configuration and the efficiency of LoRA underscore our contribution: by combining a machine-expert seed construction with an optimized bootstrapping pipeline, we effectively mitigate the domain shift caused by complex backgrounds and irregular styles without relying on massive manual annotations.

Appendix C, specifically Tables 4 and 5, reports the same grouped ablation under NES and EM. These results support the main trends while making metric-specific differences explicit.

Following the previous analysis, we provide a more granular examination of the four dimensions of the TCPI to further illustrate the domain-specific performance of our pipeline.

Clarity. Results show a trade-off between standard legibility and extreme degradation. W/o Rotate achieves the best performance on **Legible** samples (0.827), suggesting that canonical orientation is important when stroke details are discernible. Conversely, LoRA leads in the **Degraded** category (0.871), indicating its strength in handling low-quality or faded text.

Calligraphy. We observe a clear "difficulty-specialization" pattern. W/o Rotate and LoRA tie on **Regular** script (0.936), while LoRA performs best on the more challenging **Running** (0.852) and **Cursive** (0.753) scripts. This suggests that LoRA is particularly effective at capturing the fluid, abstract morphology of expressive brushwork.

Interference. Performance is heavily influenced by background noise. W/o Rotate performs best on **Clean** samples (0.869), while LoRA exhibits the strongest robustness to complex interference, leading on **Texture** (0.839) and **Seal** (0.826). This highlights LoRA's advantage in disentangling text from painting textures and seal overlaps.

Length. A segmented advantage is evident across text lengths. W/o Rotate excels on **Short** (0.868) and **Medium** (0.901) sequences. RL occupies a distinct niche, achieving the best **Long**-text score (0.774) despite weak overall performance, suggesting that its sequence-level reward helps mitigate error accumulation in extended transcriptions.

Clarity vs. Interference. A counter-intuitive yet vital observation from the ablation study is that performance on **Degraded** samples is often better than that on **Legible** samples within the Clarity dimension, whereas **Clean** samples intuitively perform better than samples with **Texture** or **Seal** interference. To clarify, all seed data passed human recognizability checks, meaning **Degraded** samples are visually faded but structurally intact. This slight lack of clarity acts as a natural augmentation, forcing the vision-language model to rely stronger on its robust linguistic priors and local context rather than overfitting to superficial stroke patterns. In contrast, **Interference** elements like seals or textures explicitly inject destructive pixel noise that overwrites the structural topology of the characters, misleading the OCR encoder. Consequently, models can overcome fading (Clarity-Degraded) through language reasoning much better than they can reconstruct corrupted visual topologies (Interference).

The ablation study clarifies the distinct roles of each strategy:

- **W/o Rotate** is the strongest general-purpose configuration by F1 (0.851), particularly for legible and regular samples.
- **LoRA** is the optimal choice for "hard cases"—specializing in degraded text, running calligraphy, and high-interference backgrounds.
- **RL** provides a specialized mechanism for long-range sequence coherence.

5 Discussion and Conclusion

The benchmark, grouped ablations, and training dynamics show that a small human-curated seed set supplies stable, low-noise supervision, while filtered pseudo-labels expand coverage of the real inscription distribution. Gains are monotonic as the expansion set grows from one fifth to one half and then the full set. W/o Rotate and LoRA provide the strongest overall performance, whereas RL has a narrower advantage on long sequences. The qualitative results in Appendix E further show that domain bootstrapping helps preserve long running-script sequences and separate expressive characters from texture and seal interference.

Transferable pipeline insights. Three lessons extend beyond TCPIs. First, data should be split by source artifact rather than by crop, preventing correlated texture and style cues from leaking across stages. Second, a verified seed set can anchor confidence-based pseudo-label expansion, making target-domain coverage scalable without requiring exhaustive manual transcription. Third, optimization should reflect the target sub-distribution: FPFT or LoRA provides strong general adaptation, while sequence-level RL can complement it when long-range coherence matters. These observations motivate iterative data expansion and staged strategy composition, rather than relying indefinitely on one fixed training scheme.

In conclusion, our five-module pipeline combines painting-level division, machine-expert seed construction, TCPI adaptation, filtered pseudo-label expansion, and bootstrapped retraining. The final model reaches an overall $F1$ of 0.851 and clearly outperforms general OCR and multimodal baselines. Rather than proposing a new OCR architecture, this work contributes a scalable and reproducible data-centric adaptation pathway for low-resource cultural heritage recognition. The independent audit in Appendix D also makes the remaining pseudo-label noise explicit, providing a basis for future uncertainty-aware filtering and iterative re-labeling.

AI Use Statement

In this work, we used generative AI tools to assist with dataset categorization, the development of model-evaluation code, the search for potentially relevant references, and language polishing of the manuscript. We did not use general-purpose generative AI assistants to formulate hypotheses or mathematical claims, design the methodology or experiments, determine final expert-verified transcription annotations, develop or run model training, run benchmark evaluations, or interpret the experimental results. The OCR models that proposed seed transcriptions and expansion pseudo-labels are part of the experimental pipeline described in Section 3; experts selected or corrected seed annotations and audited a sample of expansion labels. The authors reviewed the AI-assisted dataset categorizations, inspected and tested the evaluation code, independently verified the relevance and bibliographic accuracy of all cited references, and reviewed all language edits. We take full responsibility for the final content of this work, including all text, claims, results, code, and other artifacts produced with the aid of generative AI.

References

Nicolas Angleraud, Antonia Karamolegkou, Benoît Sagot, and Thibault Clérice. Structure-aware text recognition for ancient greek critical editions. *arXiv preprint arXiv:2603.02803*, 2026.

Leonardo Bassanini, Ludovico Biancardi, Alfio Ferrara, Andrea Gamberini, Sergio Picascia, and Folco Vaglienti. Quid est veritas? a modular framework for archival document analysis. *arXiv preprint arXiv:2603.28108*, 2026.

Kenza Benkirane, Dan Goldwater, Martin Asenov, and Aneiss Ghodsi. Disco: Document intelligence suite for comparative evaluation. *arXiv preprint arXiv:2603.23511*, 2026.

Fitsum Sileshi Beyene and Christopher L. Dancy. A survey of ocr evaluation methods and metrics and the invisibility of historical documents. *arXiv preprint arXiv:2603.25761*, 2026.

Haoyu Cao, Kun Yin, Yunfei Wu, Bing Liu, Zhongpeng Cai, Xiaotian Li, Huang Chen, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, Yunsheng Wu, Qianyu Li, Antai Guo, Yanzhen Liao, Yanqiu Qu, Haodong Lin, Chengxu He, and Shuangyin Liu. Youtu-parsing: Perception, structuring and recognition via high-parallelism decoding. *arXiv preprint arXiv:2601.20430*, 2026.

Bapu D Chendage and Rajivkumar S Mente. Integrated framework for selecting and enhancing ancient marathi inscription images from stone, metal plate, and paper documents. *arXiv preprint arXiv:2601.04800*, 2026.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl-1.5: Towards a multi-task 0.9 b vlm for robust in-the-wild document parsing. *arXiv preprint arXiv:2601.21957*, 2026a.

Cheng Cui, Yubo Zhang, Ting Sun, Xueqing Wang, Hongen Liu, Manhui Lin, Yue Zhang, Tingquan Gao, Changda Zhou, Jiaxuan Liu, et al. Pp-ocrv5: A specialized 5m-parameter model rivaling billion-parameter vision-language models on ocr tasks. *arXiv preprint arXiv:2603.24373*, 2026b.

Luigi Curini, Alfio Ferrara, Giovanni Pagano, and Sergio Picascia. Transcription and recognition of italian parliamentary speeches using vision-language models. *arXiv preprint arXiv:2603.28103*, 2026.

Adhemar de Senneville, Xavier Bou, Jérémy Anger, Rafael Grompone, and Gabriele Facciolo. Unlocking few-shot capabilities in lvlms via prompt conditioning and head selection. *arXiv preprint arXiv:2603.24181*, 2026.

Arundhathi Dev and Justin Zhan. Efficient domain adaptation for text line recognition via decoupled language models. *arXiv preprint arXiv:2603.28028*, 2026.

Daxiang Dong, Mingming Zheng, Dong Xu, Chunhua Luo, Bairong Zhuang, Yuxuan Li, Ruoyun He, Haoran Wang, Wenyu Zhang, Wenbo Wang, et al. Qianfan-ocr: A unified end-to-end model for document intelligence. *arXiv preprint arXiv:2603.13398*, 2026a.

Hejun Dong, Junbo Niu, Bin Wang, Weijun Zeng, Wentao Zhang, and Conghui He. Mineru-diffusion: Rethinking document ocr as inverse rendering via diffusion decoding. *arXiv preprint arXiv:2603.22458*, 2026b.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, et al. Glm-ocr technical report. *arXiv preprint arXiv:2603.10910*, 2026.

Hao Feng, Wei Shi, Ke Zhang, Xiang Fei, Lei Liao, Dingkang Yang, Yongkun Du, Xuecheng Wu, Jingqun Tang, Yang Liu, et al. Dolphin-v2: Universal document parsing via scalable anchor prompting. *arXiv preprint arXiv:2602.05384*, 2026.

Henry Gagnier, Sophie Gagnier, and Ashwin Kirubakaran. Kazakhocr: A synthetic benchmark for evaluating multimodal models in low-resource kazakh script ocr. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pp. 60–65, 2026.

Weile Gong, Yiping Zuo, Zijian Lu, Xin He, Weibei Fan, and Chen Dai. From plausibility to verifiability: Risk-controlled generative ocr for vision-language models. *arXiv preprint arXiv:2603.19790*, 2026.

Nayeb Hasin, Md Arafath Rahman Nishat, Mainul Islam, Khandakar Shakib Al Hasan, and Asif Newaz. A robust deep learning framework for bangla license plate recognition using yolo and vision-language ocr. *arXiv preprint arXiv:2603.10267*, 2026.

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

Nevidu Jayatilleke, Nisansa de Silva, Uthpala Nimanthi, Gagani Kulathilaka, Azra Safrullah, and Johan Sofalas. Sidiac-v.2.0: Sinhala diachronic corpus version 2.0. *arXiv preprint arXiv:2603.10861*, 2026.

Christopher Kermorvant, Eva Bardou, Manon Blanco, and Bastien Abadie. Callico: a versatile open-source document image annotation platform. In *International Conference on Document Analysis and Recognition*, pp. 338–353. Springer, 2024.

Marry Kong, Rina Buoy, Sovisal Chenda, Nguonly Taing, Masakazu Iwamura, and Koichi Kise. Towards universal khmer text recognition. *arXiv preprint arXiv:2603.00702*, 2026.

Minduli Lasandi and Nevidu Jayatilleke. Sinhalegal: A benchmark corpus for information extraction and analysis in sinhala legislative texts. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pp. 114–131, 2026.

Anh Le and Asanobu Kitamoto. Training kindai ocr with parallel textline images and self-attention feature distance-based loss. *arXiv preprint arXiv:2508.08537*, 2025.

Gengluo Li, Xingyu Wan, Shangpin Peng, Weinong Wang, Hao Feng, Yongkun Du, Binghong Wu, Zheng Ruan, Zhiqiong Lu, Liang Wu, Pengyuan Lyu, Huawen Shen, Zibin Lin, Shijing Hu, Jieneng Yang, Hongbing Wen, Guanghua Yu, Hong Liu, Bochao Wang, Can Ma, Han Hu, Chengquan Zhang, and Yu Zhou. Hunyuanocr-1.5: Making lightweight ocr vlms faster and better. *arXiv preprint arXiv:2607.04884*, 2026.

Xin Li, Mingming Gong, Yunfei Wu, Jianxin Dai, Antai Guo, Xinghua Jiang, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. Dream: Document reconstruction via end-to-end autoregressive model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2949–2957, 2025a.

Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. dots. ocr: Multilingual document layout parsing in a single vision-language model. *arXiv preprint arXiv:2512.02498*, 2025b.

Tingyu Lin, Marco Peer, Florian Kleber, and Robert Sablatnig. Clappertext: A benchmark for text recognition in low-resource archival. In *Document Analysis and Recognition–ICDAR 2025 Workshops: Wuhan, China, September 20–21, 2025, Proceedings, Part II*, pp. 329. Springer Nature, 2026.

Bonan Liu, Zeyu Zhang, Bingbing Meng, Han Wang, Hanshuo Zhang, Chengping Wang, Daji Ergu, and Ying Cai. Omniocr: Generalist ocr for ethnic minority languages. *arXiv preprint arXiv:2602.21042*, 2026.

Dongliang Luo, Hanshen Zhu, Ziyang Zhang, Dingkang Liang, Xudong Xie, Yuliang Liu, and Xiang Bai. Semiets: Integrating spatial and content consistencies for semi-supervised end-to-end text spotting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9329–9338, 2025.

Koki Maeda and Naoaki Okazaki. Jawildtext: A benchmark for vision-language models on japanese scene text understanding. *arXiv preprint arXiv:2603.27942*, 2026.

Erez Meoded. Handwritten text recognition of historical manuscripts using transformer-based models. *arXiv preprint arXiv:2508.11499*, 2025.

H Neji, J Nogueras-Iso, J Lacasta, MÁ Latre, and FJ García-Marco. Fp-thd: Full page transcription of historical documents. *arXiv preprint arXiv:2601.17040*, 2026.

Hieu Minh Nguyen, Tam Le-Thanh Dang, and Kiet Van Nguyen. Towards signboard-oriented visual question answering: Visignvqa dataset, method and benchmark. *arXiv preprint arXiv:2512.22218*, 2025.

Vik Paruchuri. chandra: Ocr model that handles complex tables, forms, handwriting with full layout. GitHub repository, 2025. URL `https://github.com/datalab-to/chandra/releases/tag/v0.1.0`.

Jake Poznanski, Luca Soldaini, and Kyle Lo. olmocr 2: Unit test rewards for document ocr. *arXiv preprint arXiv:2510.19817*, 2025.

Sina Semnani, Han Zhang, Xinyan He, Merve Tekgürler, and Monica Lam. 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*, pp. 34765–34812, 2025.

Guowei Tang. Data organization matters in multimodal instruction tuning: A controlled study of capability trade-offs. *arXiv preprint arXiv:2603.27744*, 2026.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, et al. Hunyuanocr technical report. *arXiv preprint arXiv:2511.19575*, 2025.

Selim Furkan Tekin, Yichang Xu, Gaowen Liu, Ramana Rao Kompella, Margaret L Loper, and Ling Liu. Vision verification enhanced fusion of vlms for efficient visual reasoning. *arXiv preprint arXiv:2603.12669*, 2026.

Phat Tran, Phuoc Pham, Hung Trinh, and Tho Quan. Low-resource heuristics for bahnaric optical character recognition improvement. *arXiv preprint arXiv:2601.02965*, 2026.

Ari Vesalainen, Eetu Mäkelä, Laura Ruotsalainen, and Mikko Tolonen. Error patterns in historical ocr: A comparative analysis of trocr and a vision-language model. *arXiv preprint arXiv:2602.14524*, 2026.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, Bangrui Xu, Junbo Niu, Mengzhang Cai, Jiantao Qiu, Qintong Zhang, Dongsheng Ma, Yuefeng Sun, Hejun Dong, Wenzheng Zhang, Jutao Xiao, Jiayong Shi, Pengyu Liao, Xiaomeng Zhao, Huaping Zhong, Liqun Wei, Jing Yu, Jie Yang, Wei Li, Shasha Wang, Qianqian Wu, Xuanhe Zhou, Weijia Li, Zhenxiang Li, Zhongying Tu, Jiang Wu, Lijun Wu, Chao Xu, Kai Chen, Wentao Zhang, Yu Qiao, Bowen Zhou, Dahua Lin, and Conghui He. Mineru2.5-pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*, 2026.

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026.

Petros Georgoulas Wraight, Giorgos Sfikas, Ioannis Kordonis, Petros Maragos, and George Retsinas. Optimal transport for handwritten text recognition in a low-resource regime. *arXiv preprint arXiv:2509.16977*, 2025.

Hao Wu, Haoran Lou, Xinyue Li, Zuodong Zhong, Zhaojun Sun, Phellon Chen, Xuanhe Zhou, Kai Zuo, Yibo Chen, Xu Tang, et al. Firered-ocr technical report. *arXiv preprint arXiv:2603.01840*, 2026.

Wenwen Yu, Zhibo Yang, Jianqiang Wan, Sibo Song, Jun Tang, Wenqing Cheng, Yuliang Liu, and Xiang Bai. Omniparser v2: Structured-points-of-thought for unified visual text parsing and its generality to multimodal large language models. *arXiv preprint arXiv:2502.16161*, 2025.

Qintong Zhang, Junyuan Zhang, Zhifei Ren, Linke Ouyang, Zichen Wen, Junbo Niu, Yuan Qu, Bin Wang, Ka-Ho Chow, Conghui He, et al. Docr-inspector: Fine-grained and automated evaluation of document parsing with vlm. *arXiv preprint arXiv:2512.10619*, 2025.

Zelun Zhang, Hongen Liu, Suyin Liang, Yubo Zhang, Yiqing Xiang, Jiaxuan Liu, Ting Sun, Manhui Lin, Yue Zhang, Changda Zhou, Tingquan Gao, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr-vl-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. *arXiv preprint arXiv:2606.03264*, 2026.

Yixin Zhao, Yuyi Zhang, and Lianwen Jin. Mccd: A multi-attribute chinese calligraphy character dataset annotated with script styles, dynasties, and calligraphers. In *International Conference on Document Analysis and Recognition*, pp. 538–555. Springer, 2025.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*, 2024.

A Representative TCPI Example

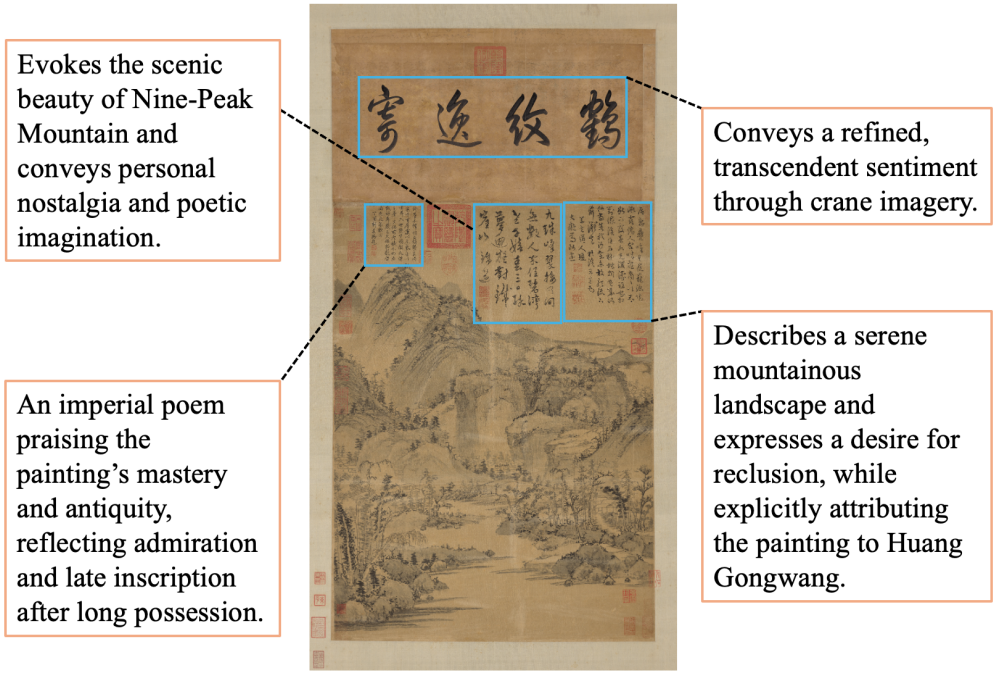

Figure 4: Nine-Peak Emerald, painted by Huang Gongwang (1269–1354), serves as a representative TCPI example. It contains four inscriptions by different authors that collectively convey key information about the TCP.

B Reward Function

This appendix provides a detailed specification of the reward function used for reinforcement learning. The reward is designed for structured optical character recognition (OCR), where each model response consists of two components: (i) a YAML front matter containing document-level metadata and (ii) the recognized textual content. Accordingly, the reward jointly evaluates structural compliance, textual fidelity, and response length.

Let x denote an input document image, y denote a response sampled from the policy model, and $y^\star$ denote the corresponding reference response. We decompose each response into a YAML block and a textual body:

$$y = \big(Y(y), T(y)\big), \qquad y^\star = \big(Y(y^\star), T(y^\star)\big), \tag{1}$$

where $Y(\cdot)$ extracts the YAML front matter and $T(\cdot)$ extracts the subsequent OCR text. The scalar reward assigned to y is composed of a format-compliance reward, a character-level recall reward, and a soft overlength penalty.

Format-compliance reward. The YAML front matter is expected to contain a predefined set of metadata fields. Let $\mathcal{K}(Y)$ denote the set of top-level keys appearing in a YAML block Y. We define the format reward as

$$R_{\text{fmt}}(y, y^\star) = \mathbb{I}\left[\mathcal{K}(Y(y^\star)) \subseteq \mathcal{K}(Y(y))\right], \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Thus, $R_{\text{fmt}} = 1$ if the generated response contains all required YAML keys, and $R_{\text{fmt}} = 0$ otherwise. This definition permits additional fields in the generated YAML block and therefore evaluates structural completeness rather than exact string equality.

When the reference response does not explicitly provide a set of YAML fields, we use the following default required-key set:

$$\mathcal{K}_{\text{default}} = \{\texttt{is_text_clear}, \texttt{primary_language}, \texttt{is_rotation_valid}, \texttt{rotation_correction}, \texttt{is_table}, \texttt{is_diagram}\}. \tag{3}$$

The format reward only checks the presence of these fields. Their semantic correctness is not independently scored by this term.

Character-level recall reward. Before measuring textual fidelity, both the generated text and the reference text are normalized. Specifically, traditional Chinese characters are converted to simplified Chinese, and all whitespace characters are removed:

$$\widetilde{T}(y) = \text{RemoveSpace}\left(\text{T2S}\left(T(y)\right)\right), \tag{4}$$

where $\text{T2S}(\cdot)$ denotes traditional-to-simplified Chinese conversion. The same transformation is applied to $T(y^\star)$.

Let $\mathcal{M}\big(\widetilde{T}(y), \widetilde{T}(y^\star)\big)$ denote the collection of non-overlapping, order-preserving character spans identified as equal by sequence alignment. For a matched span m, let $|m|$ denote its number of characters. The textual reward is defined as the fraction of reference characters covered by the aligned equal spans:

$$R_{\text{rec}}(y, y^\star) = \frac{\sum_{m \in \mathcal{M}\left(\widetilde{T}(y), \widetilde{T}(y^\star)\right)} |m|}{\left|\widetilde{T}(y^\star)\right|}. \tag{5}$$

This recall-oriented definition emphasizes preserving the textual information contained in the reference. Missing reference characters directly decrease the score, whereas unmatched insertions do not directly increase the denominator. Nevertheless, insertions may affect the sequence alignment and thereby reduce the total length of aligned equal spans.

For completeness, when the normalized reference text is empty, we define $R_{\text{rec}} = 1$ only if the generated text is also empty; otherwise, $R_{\text{rec}} = 0$. For a non-empty reference, $R_{\text{rec}} \in [0, 1]$.

Structured OCR reward. We combine structural compliance and textual fidelity into a structured OCR reward:

$$R_{\text{OCR}}(y, y^\star) = 0.2 R_{\text{fmt}}(y, y^\star) + 0.8 R_{\text{rec}}(y, y^\star). \tag{6}$$

The larger weight on R_{rec} reflects the primary objective of recovering the document content, while the format term encourages the model to produce machine-readable metadata. Since both constituent scores are bounded in $[0, 1]$, the resulting OCR reward also satisfies $R_{\text{OCR}} \in [0, 1]$.

Soft overlength penalty. Long responses substantially increase rollout and optimization costs and may contain redundant or repetitive content. We therefore introduce a soft overlength penalty based on the number of generated tokens. Let $L(y)$ denote the token length of response y, $L_{\max}$ denote the maximum generation length, and L_{cache} denote the length of the penalty interval. The penalty is

$$R_{\text{len}}(y) = \min\left(-\frac{L(y) - (L_{\max} - L_{\text{cache}})}{L_{\text{cache}}},\, 0\right), \tag{7}$$

Equivalently, no penalty is applied when $L(y) \le L_{\max} - L_{\text{cache}}$. The reward then decreases linearly from 0 to -1 as the response length increases from $L_{\max} - L_{\text{cache}}$ to $L_{\max}$ tokens:

$$R_{\text{len}}(y) = \begin{cases} 0, & L(y) \le L_{\max} - L_{\text{cache}}, \\ -\dfrac{L(y) - (L_{\max} - L_{\text{cache}})}{L_{\text{cache}}}, & L_{\max} - L_{\text{cache}} < L(y) \le L_{\max}. \end{cases} \tag{8}$$

This soft penalty provides a gradual learning signal before the hard generation limit is reached.

Final reward. The final reward used for group-relative advantage estimation is

$$\boxed{R(y, y^\star) = 0.3R_{\text{len}}(y) + 0.7R_{\text{OCR}}(y, y^\star)}. \tag{9}$$

Substituting Eq. (6) into Eq. (9) gives

$$\boxed{R(y, y^\star) = 0.3R_{\text{len}}(y) + 0.14R_{\text{fmt}}(y, y^\star) + 0.56R_{\text{rec}}(y, y^\star)}. \tag{10}$$

Consequently, the reward prioritizes character recovery while retaining an explicit incentive for structural validity. The length term acts only as a non-positive regularizer and therefore cannot compensate for inaccurate OCR content. Under the configured generation limit, the final reward lies in $[-0.3, 0.7]$, where the upper bound is attained by a structurally valid, textually correct, and sufficiently concise response.

Notation summary.

- x is the input document image.
- y is the generated response.
- $y^\star$ is the reference response.
- $Y(\cdot)$ extracts the YAML front matter.
- $T(\cdot)$ extracts the OCR body text.
- $\mathcal{K}(\cdot)$ extracts top-level YAML keys.
- $\mathbb{I}[\cdot]$ is an indicator function.
- $\text{T2S}(\cdot)$ converts traditional Chinese to simplified Chinese.
- $\text{RemoveSpace}(\cdot)$ removes whitespace.
- $\mathcal{M}(\cdot, \cdot)$ is the set of aligned equal character spans.
- $|m|$ is the length of a matched span.
- $L(y)$ is the generated token length.
- $L_{\max}$ is the maximum generation length.
- L_{cache} is the length of the soft overlength penalty interval.

C Evaluation Metric Definitions

Let $\hat{s}_i$ and s_i denote the normalized prediction and reference for sample i, respectively. Normalization converts traditional Chinese to simplified Chinese, extracts the OCR body, and removes formatting artifacts. Let M_i be the total number of characters in the non-overlapping, order-preserving aligned equal spans between $\hat{s}_i$ and s_i. The sample-level character precision, recall, and $F1$ are

$$p_i = \frac{M_i}{|\hat{s}_i|}, \qquad r_i = \frac{M_i}{|s_i|}, \qquad f_i = \frac{2p_i r_i}{p_i + r_i}. \tag{11}$$

Normalized edit similarity and exact match are defined as

$$\text{NES}_i = 1 - \frac{d_{\text{lev}}(\hat{s}_i, s_i)}{\max(|\hat{s}_i|, |s_i|)}, \qquad \text{EM}_i = \mathbb{I}[\hat{s}_i = s_i], \tag{12}$$

where d_{lev} is the character-level Levenshtein distance. When both normalized strings are empty, all five scores are defined as one; when exactly one is empty, they are defined as zero. For an evaluation set of N samples, the reported metrics are the corresponding sample averages:

$$\begin{aligned} P &= \frac{1}{N}\sum_{i=1}^{N} p_i, \qquad R = \frac{1}{N}\sum_{i=1}^{N} r_i, \qquad F1 = \frac{1}{N}\sum_{i=1}^{N} f_i, \\ \text{NES} &= \frac{1}{N}\sum_{i=1}^{N} \text{NES}_i, \quad \text{EM} = \frac{1}{N}\sum_{i=1}^{N} \text{EM}_i. \end{aligned} \tag{13}$$

Table 4: Fine-grained evaluation on the seed dataset (NES score).

Settings	Overall	Clarity		Calligraphy			Interference			Length		
		Degraded	Legible	Running	Cursive	Regular	Clean	Texture	Seal	Short	Medium	Long
Baseline	0.823	0.852	0.784	0.823	0.727	0.911	0.847	0.800	0.786	0.840	0.880	0.728
1/2 Samples	0.815	0.834	0.791	0.815	0.691	0.917	0.834	0.797	0.785	0.835	0.849	0.735
1/5 Samples	0.796	0.818	0.766	0.793	0.678	0.918	0.824	0.771	0.752	0.811	0.833	0.725
Only Rotate	0.796	0.825	0.758	0.793	0.680	0.904	0.823	0.778	0.753	0.814	0.839	0.711
W/o Rotate	**0.829**	0.854	**0.795**	**0.829**	0.726	0.917	**0.848**	0.812	**0.798**	**0.851**	**0.885**	0.724
LoRA	0.827	**0.857**	0.788	**0.829**	**0.729**	0.917	0.844	**0.818**	0.797	0.847	0.879	0.731
RL	0.787	0.806	0.761	0.783	0.667	**0.919**	0.813	0.765	0.747	0.785	0.836	**0.750**

Table 5: Fine-grained evaluation on the seed dataset (EM score).

Settings	Overall	Clarity		Calligraphy			Interference			Length		
		Degraded	Legible	Running	Cursive	Regular	Clean	Texture	Seal	Short	Medium	Long
Baseline	**0.332**	**0.349**	**0.309**	**0.312**	0.128	**0.479**	**0.401**	0.231	0.196	**0.502**	0.188	**0.014**
1/2 Samples	0.268	0.292	0.235	0.249	0.093	0.388	0.343	0.170	0.114	0.436	0.061	0.005
1/5 Samples	0.243	0.262	0.218	0.222	0.078	0.372	0.320	0.153	0.090	0.404	0.039	0.000
Only Rotate	0.270	0.296	0.235	0.250	0.069	0.397	0.347	0.170	0.120	0.429	0.083	**0.014**
W/o Rotate	0.329	0.346	0.307	0.309	**0.142**	0.463	0.396	0.228	0.196	0.500	0.177	**0.014**
LoRA	0.327	0.346	0.302	0.309	**0.142**	0.455	0.379	**0.242**	**0.209**	0.493	**0.199**	0.005
RL	0.228	0.245	0.205	0.201	0.049	0.397	0.301	0.129	0.087	0.369	0.061	0.005

Ablation robustness under complementary metrics. The main paper reports the fine-grained ablation results using character-level $F1$. Tables 4 and 5 provide the corresponding NES and EM results. The principal trends are consistent across the three metrics. Increasing the pseudo-labeled training set from 1/5 to 1/2 and then to the full set improves the overall score monotonically. Only Rotate is consistently weaker than the Baseline, whereas W/o Rotate and LoRA remain the strongest general-purpose configurations. RL has lower overall scores but retains an advantage on long text under $F1$ and NES. The strict EM metric does not reproduce this long-text advantage because even one incorrect character causes the complete prediction to be counted as a mismatch. EM also ranks the Baseline marginally above W/o Rotate overall, while $F1$ and NES rank W/o Rotate first. These local differences reflect the metrics' distinct sensitivities rather than a change in the main conclusions.

D Human-Expert Quality Validation of the Expansion Dataset

To independently verify the quality of the pseudo-label expansion dataset, we conducted a human audit on samples drawn from the same distribution as the full expansion dataset. Human experts transcribed the sampled inscription images to establish human-verified references. These expert annotations were then compared with the pseudo-labels produced by the TCPI-adapted model. The resulting audit set contains 463 samples, covering 9.37% of the 4,943-sample expansion dataset. This protocol directly measures pseudo-label fidelity on the data used for bootstrapped training, rather than inferring its quality from downstream model performance or from the rule-based filtering score alone.

We use the same character-level P, R, $F1$, NES, and EM metrics defined in Section C. Before evaluation, both expert references and pseudo-labels are converted from traditional to simplified Chinese; OCR formatting artifacts, punctuation, whitespace, and control characters are removed. All values in Table 6 are macro averages over samples. To examine whether transcription length affects pseudo-label quality, we additionally partition the audit set by the expert-reference OCR-body length using the same boundaries as the main paper: short (0–50 characters), medium (51–100 characters), and long (more than 100 characters).

Table 6: Expert audit of pseudo-label quality on 463 distribution-matched samples from the expansion dataset. The pseudo-labels are evaluated against human-expert annotations.

Split	Samples	P	R	$F1$	NES	EM
Overall	463	0.823	0.802	0.809	0.785	0.095
Short (0–50)	251	0.817	0.809	0.812	0.793	0.163
Medium (51–100)	115	0.822	0.800	0.808	0.787	0.026
Long (>100)	97	0.840	0.784	0.802	0.763	0.000

The audit yields an overall character-level $F1$ of 0.809 and NES of 0.785. Performance is stable across the three length groups: $F1$ ranges only from 0.802 to 0.812. For long inscriptions, precision increases to 0.840 while recall decreases to 0.784, indicating that omissions, rather than unsupported insertions, become the more prominent error mode as sequence length grows. EM declines sharply with length and reaches zero for long inscriptions because it requires every character in an extended sequence to match exactly; therefore, the character-level $F1$ and NES scores provide a more informative assessment of partial transcription fidelity for this audit.

Table 7 further reports cumulative sample-level agreement at several thresholds. In particular, 65.01% of the audited pseudo-labels achieve $F1 \geq 0.80$, and 91.58% achieve $F1 \geq 0.50$. The corresponding proportions under NES are 60.91% and 88.98%, respectively. These independently verified results provide direct evidence that the filtered expansion set contains substantial usable character-level supervision. They also reveal a residual low-agreement tail, making the audit a more conservative and transparent quality check than the automatic filtering score alone.

Table 7: Cumulative sample-level agreement in the expert audit. Each row reports the number and proportion of audited pseudo-labels whose score is at least the specified threshold.

Metric	Threshold	Samples	Proportion
$F1$	≥ 0.95	111	23.97%
$F1$	≥ 0.90	187	40.39%
$F1$	≥ 0.80	301	65.01%
$F1$	≥ 0.50	424	91.58%
NES	≥ 0.95	105	22.68%
NES	≥ 0.90	175	37.80%
NES	≥ 0.80	282	60.91%
NES	≥ 0.50	412	88.98%

Although the audited expansion pseudo-labels have lower $F1$, NES, and EM scores than the bootstrapped Baseline on the seed benchmark, the two results come from different image sets and do not establish a direct label-to-model improvement; together with the gains from larger pseudo-labeled training sets in Table 3, they suggest that the pipeline can learn effectively despite residual pseudo-label noise.

E Qualitative Analysis

To intuitively illustrate the performance advantages of the proposed TCPI-bootstrapped model, we provide a qualitative comparison against the base model (olmOCR-2-7B-1025) and a general-purpose LLM (GPT-5.4) under challenging conditions, as shown in Figure 5. Specifically, the results of our proposed TCPI-bootstrapped model, the base model, and GPT-5.4 are presented as **Model1**, **Model2**, and **Model3**, respectively.

Sample 1 (Left) represents a Long, Degraded, and Running script inscription with Clean background semantics. While GPT-5.4 struggles with sequence coherence over extended texts and the base olmOCR model occasionally guesses semantics rather than strictly following the visual evidence, our TCPI-bootstrapped model effectively maintains linguistic consistency and accurately predicts the fluid morphological transitions of the expressive running script.

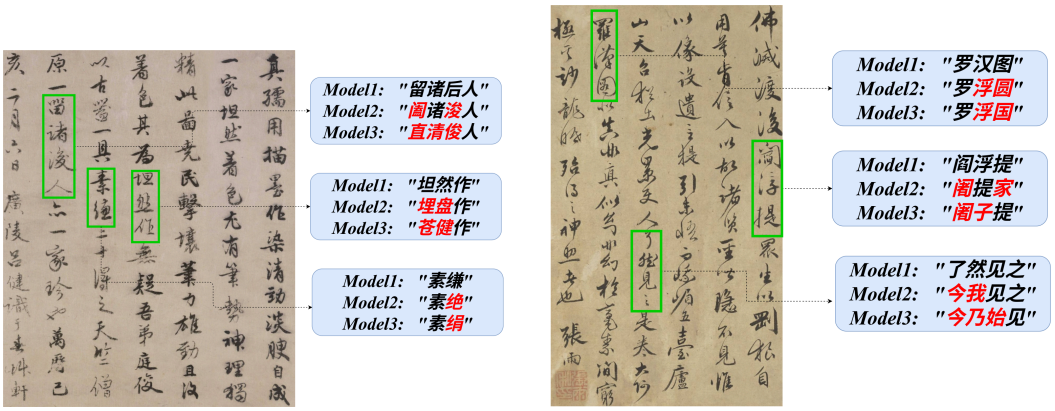

Figure 5: Qualitative comparison of recognition results. Left (Sample 1): Long, Degraded, Running script, and Clean. Right (Sample 2): Medium, Degraded, mixed (Running and Cursive) script, Texture and Seal interference. Red text in the figures indicates incorrect model predictions.

Sample 2 (Right) features a Medium-length, Degraded inscription containing mixed calligraphy (both Running and Cursive scripts) under heavy painting interference. General LLMs like GPT-5.4 are severely disrupted by painting-texture artifacts, leading to structural hallucinations or omitted tokens. The base olmOCR model frequently fails to disentangle the highly abstract cursive strokes from the intense background textures. In contrast, our TCPI-bootstrapped model—enhanced by domain-specific training on high-quality pseudo-labels—demonstrates robust disentanglement capabilities, faithfully extracting the authentic text despite the heavy visual noise and expressive brushwork.

These qualitative examples reinforce our quantitative findings: while general VLMs struggle with the artistic variability and low-resource nature of TCPIs, our domain-centric bootstrapping pathway successfully internalizes the complex visual and structural priors required for heritage OCR. By consistently outperforming general-purpose models in handling intricate interference and styled calligraphy, our pipeline naturally yields actionable lessons for other similarly constrained domains.