ONE GLYPH, MANY TOKENS: TOKENIZATION BOTTLENECKS IN GENERATIVE OCR

Anonymous authors
Paper under double-blind review

ABSTRACT

In generative optical character recognition (OCR), a visually atomic character may correspond to multiple output tokens. This representation mismatch forces the model to transcribe a single glyph through a sequence of autoregressive predictions, increasing the risk of error propagation. Our analysis reveals a pronounced degradation in OCR performance as token fragmentation increases, suggesting that output representation itself can remain a bottleneck. To address this, we introduce Glyph-Aligned Output-Space Adaptation (GAOSA), a representation-level adaptation approach that aligns the output granularity of fragmented characters with visually atomic glyphs. GAOSA replaces fragmented multi-token decoding paths with atomic character-level output units. It initializes these units from their constituent-token representations and jointly adapts the token embeddings and language-modeling head under transcription supervision. We primarily study GAOSA in historical Chinese OCR, where standard SFT improves target-character F1 from 7.6% to 59.7%, while GAOSA further raises it to 88.6%, demonstrating substantial gains beyond those achieved through supervision alone. To support full-page transcription, we further synthesize target characters into historical layouts and train with page-level supervision, enabling the model to jointly learn layout understanding and character recognition. The resulting model achieves a normalized edit distance (NED) of 0.11 on AncientDoc, compared with 0.19 for PaddleOCR-VL-1.6 and 0.18 for Qwen3.6-27B. In addition, GAOSA yields consistent gains across distinct scripts and symbol systems. On HHD-Ethiopic, GAOSA improves F1 from 42.6% (SFT only) to 76.2%; on OCR-IPA, it improves F1 from 73.9% to 90.8%. These results establish glyph-aligned output representation as an effective intervention beyond conventional SFT across diverse scripts and symbol systems. The model and benchmarks will be publicly available.

1 INTRODUCTION

Generative optical character recognition (OCR) formulates document reading as image-conditioned sequence generation (Kim et al., 2022a; Wei et al., 2024). Yet visual and output granularity need not align: a visually atomic glyph may be decomposed into multiple tokens by the pretrained tokenizer. We observe this fragmentation in historical Chinese, Ethiopic, and International Phonetic Alphabet (IPA), where one glyph may require several autoregressive predictions. This raises a fundamental question: *to what extent does representing a visually atomic character as multiple output tokens itself become a bottleneck for OCR?*

Tokenization can substantially affect multilingual language modeling (Rust et al., 2021), but in generative OCR it creates a direct mismatch between the visual target and the autoregressive prediction unit. A character fragmented into k tokens requires k successive predictions, and an early error can alter subsequent prefix conditioning. SFT can improve these predictions but does not change the underlying output granularity, leaving fragmentation as a potential residual bottleneck. We therefore treat character-level output representation as an explicit component of generative OCR.

We first examine this phenomenon in historical Chinese OCR across VLMs. Figure 1(a) reports pretrained Qwen3.5-4B results on *Qimin Yaoshu* characters grouped by token count. While 59% are single-token characters, 41% require multiple tokens. Exact isolated-character accuracy drops

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.

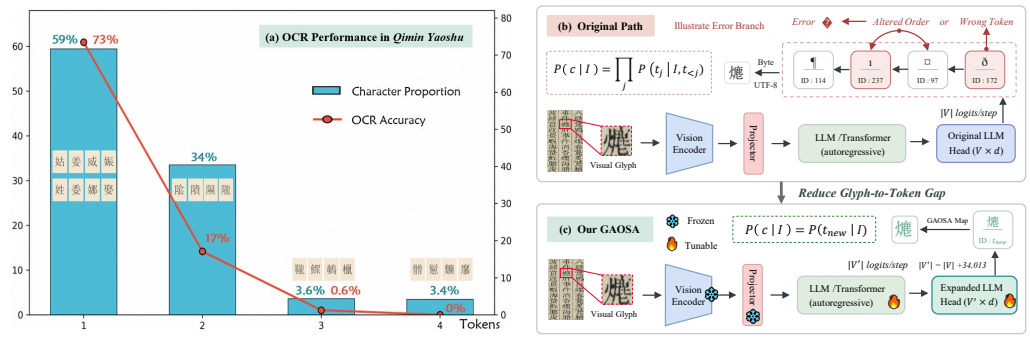

Figure 1: Token fragmentation in generative OCR and our GAOSA approach. (a) Recognition performance decreases sharply as token fragmentation increases in Chinese historical document (*Qimin Yaoshu*). (b) A visually atomic character is generated through a fragmented multi-token decoding path. (c) GAOSA aligns output granularity with the visual glyph by introducing an atomic character-level output unit and adapting the corresponding output representations.

from 73% at one token to 17%, 0.6%, and 0% at two, three, and four tokens. InternVL3.5-4B shows the same degradation. Because character frequency and training exposure may also affect difficulty, we do not treat these observations alone as causal evidence. Instead, they motivate a controlled question: *under matched supervision, does aligning output granularity with the visual character resolve errors that conventional SFT leaves behind?*

To answer this question, we introduce Glyph-Aligned Output-Space Adaptation (GAOSA), a representation-level adaptation approach that aligns fragmented character outputs with visually atomic glyphs. GAOSA represents each selected fragmented character as an atomic character-level output unit instead of a multi-token decoding path, as illustrated in Figure 1(b–c). These units are initialized from constituent-token representations, and GAOSA jointly adapts the token embeddings and language-modeling head under transcription supervision. The visual target remains unchanged while the output unit matches its granularity.

We instantiate GAOSA on Qwen3.5-4B (Qwen Team, 2026a) and compare it with conventional SFT using the same fine-tuning data on our fragmentation-enriched benchmark, FELDoc-1K. Data-matched Stage-2 SFT under τ_0 raises target-character F1 from 7.6% to 59.7%, while GAOSA further improves it to 88.6%, yielding a 28.9-point gain. The benefit further extends beyond historical Chinese OCR: on HHD-Ethiopic (Belay et al., 2024), GAOSA improves F1 from 42.56% with SFT to 76.17%, and on OCR-IPA (Okabe et al., 2026), from 73.90% to 90.78%. On InternVL3.5-4B (Wang et al., 2025), GAOSA reaches 93.74% F1 and shows the same fragmentation-stratified pattern. Together, these results across writing systems and backbones support output granularity as a recurring bottleneck in generative OCR.

Character-level recognition alone is insufficient for full-page transcription, which also requires reading order and complex layouts such as vertical columns and interleaved annotations (Yu et al., 2025). We therefore adopt a two-stage training framework. For scalable page-level supervision, we synthesize historical pages using LLM-generated layout coordinates, historical fonts, and aging/degradation effects. The first stage learns full-page transcription from synthetic and manually annotated pages. In the second stage, GAOSA is introduced using target-character-enriched synthetic pages together with annotated real pages, enabling glyph-aligned output adaptation within realistic page contexts. This design combines document-level layout understanding with robust recognition of token-fragmented characters.

For full-page transcription, we evaluate the resulting two-stage model on two public benchmarks, AncientDoc (Yu et al., 2025) and MTHv2 (Ma et al., 2020), and two in-house benchmarks: HCLDoc-582 for general historical-page transcription and FELDoc-1K for fragmentation-enriched evaluation. On AncientDoc, the model achieves an NED of 0.11, compared with 0.18 for Qwen3.6-27B and 0.19 for PaddleOCR-VL-1.6. On HCLDoc-582, it achieves an NED of 0.14. More importantly, on the same FELDoc-1K used for target-character evaluation, it achieves a page-level NED of 0.03 and a target-character F1 of 88.6% for fragmented characters ($k > 1$). These results

show that the model improves fragmented-character recognition while preserving strong full-page transcription.

Our main contributions are summarized as follows:

- **Diagnosing token fragmentation as an output-representation bottleneck.** We identify the granularity mismatch between visually atomic glyphs and fragmented autoregressive outputs, and show that recognition performance deteriorates sharply as fragmentation increases. Matched-supervision comparisons further reveal substantial residual errors that persist despite effective supervised adaptation.
- **Glyph-Aligned Output-Space Adaptation.** We propose GAOSA, a representation-level adaptation approach that aligns model output granularity with visually atomic glyphs. GAOSA introduces atomic character-level output units, initializes them from constituent-token representations, and jointly adapts token embeddings and the language-modeling head under transcription supervision.
- **Validation across pages, scripts, and backbones.** We integrate GAOSA into full-page historical document transcription and validate it on both public and in-house benchmarks. We further demonstrate gains over data-matched SFT on Ethiopic and IPA, together with compatibility on an additional VLM backbone, showing that the benefit of glyph-aligned output representation extends beyond a single script or model.

2 Related Work

2.1 Generative OCR

TrOCR (Li et al., 2023) formulates OCR as end-to-end image-conditioned sequence generation, while Donut (Kim et al., 2022a) extends this paradigm to OCR-free document understanding by directly generating structured text outputs. GOT (Wei et al., 2024) further broadens generative OCR toward OCR-2.0, covering text, formulas, tables, charts, and formatted outputs. General-purpose VLMs such as Qwen2.5-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025) have also become strong document readers, while specialized systems such as PaddleOCR-VL-1.6 (Zhang et al., 2026) target document parsing through dedicated data and post-training. These advances improve document-level recognition, but largely leave the tokenizer-defined output granularity unchanged.

2.2 Tokenization and Output-Space Adaptation

Tokenizer design affects multilingual modeling, sequence length, and transfer efficiency (Rust et al., 2021; Petrov et al., 2023). Existing vocabulary-adaptation methods primarily target cross-lingual transfer or generation efficiency: FOCUS (Dobler & de Melo, 2023) uses overlap-based initialization, EEVE (Kim et al., 2024) combines constituent-based initialization with staged updates, ZeTT (Minixhofer et al., 2024) predicts representations for unseen tokenizers, and Yamaguchi et al. (2026) study adaptation under limited target-language data. GAOSA instead focuses on an OCR-specific mismatch in which a visually atomic glyph is generated through several autoregressive output units. We therefore study whether changing output granularity removes errors that remain after supervised adaptation.

2.3 OCR for Specialized Character Inventories

Uncommon scripts and symbol systems motivate dedicated representations and benchmarks. Prior work includes grapheme-aware Hangul recognition (Kim et al., 2022b), sheet-music transcription (Ríos-Vila et al., 2024), historical Chinese glyph collections (Guan et al., 2024), HHD-Ethiopic (Belay et al., 2024), and multilingual OCR4MT (Ignat et al., 2022). These settings show that rare or specialized character inventories remain challenging even for modern recognizers, motivating our study of the output representation itself rather than only additional supervision.

3 METHODOLOGY

GAOSA replaces tokenizer-fragmented character representations with atomic output units while preserving page-to-sequence generation. It combines glyph-aligned output construction, constituent-anchored initialization, and contextualized page supervision in a two-stage framework; c denotes an atomic transcription character.

3.1 GLYPH–TOKEN GRANULARITY MISMATCH

Let τ_0 denote the pretrained tokenizer and Θ_0 the corresponding VLM. For a target character c, its standalone encoding is

$$\tau_0(c) = (t_1^c, \ldots, t_{k(c)}^c), \tag{1}$$

where $k(c) = |\tau_0(c)|$ is its *fragmentation degree*; $k(c) > 1$ indicates fragmentation. Standalone encoding is used only for diagnosis and subset construction, while complete transcriptions are tokenized end-to-end during training.

Given an isolated character image I_c and OCR instruction q, let $x_c = (I_c, q)$. Correct generation under τ_0 requires

$$p_{\Theta_0}(\tau_0(c) \mid x_c) = \prod_{j=1}^{k(c)} p_{\Theta_0}(t_j^c \mid x_c, t_{<j}^c). \tag{2}$$

Thus, one visual target may require several conditional predictions, and an early error can alter later prefix conditioning. Equation 2 characterizes the representation mismatch without assuming that fragmentation alone determines recognition difficulty.

3.2 GLYPH-ALIGNED OUTPUT-SPACE ADAPTATION

Atomic output-space construction. Given a target inventory $\mathcal{C}_{\text{target}}$, we select fragmented characters

$$\mathcal{C} = \{c \in \mathcal{C}_{\text{target}} : k(c) > 1\}, \tag{3}$$

and introduce one ordinary token u_c for each $c \in \mathcal{C}$. The adapted vocabulary and tokenizer satisfy

$$\mathcal{V}_1 = \mathcal{V}_0 \cup \{u_c : c \in \mathcal{C}\}, \qquad \tau_1(c) = (u_c). \tag{4}$$

Existing token IDs and special-token assignments remain unchanged. We additionally verify complete transcriptions to ensure that target characters map to their atomic units in context rather than only in standalone strings. In historical Chinese, GAOSA covers 34,013 fragmented characters; the exact tokenizer-ID allocation and physical vocabulary resizing are reported in Appendix D.

Algorithm 1. GAOSA Construction

Require: $\tau_0, \mathcal{V}_0, E_0, O_0$ and target inventory $\mathcal{C}_{\text{target}}$
Ensure: **1:** Glyph-aligned tokenizer and vocabulary $(\tau_1, \mathcal{V}_1)$
2: Constituent-anchored representations (E_1, O_1)

$\mathcal{C} \leftarrow \{c \in \mathcal{C}_{\text{target}} : |\tau_0(c)| > 1\}$
$\mathcal{V}_1 \leftarrow \mathcal{V}_0 \cup \{u_c : c \in \mathcal{C}\}$
$\tau_1 \leftarrow \tau_0$
Resize the tied input/output table to the adapted size and copy original rows
for all $c \in \mathcal{C}$ **do**
  $\mathbf{t}^c \leftarrow \tau_0(c)$
  $\tau_1(c) \leftarrow (u_c)$
  $E_1[u_c] \leftarrow \text{Mean}_{t \in \mathbf{t}^c} E_0[t]$
  $O_1[u_c] \leftarrow \text{Mean}_{t \in \mathbf{t}^c} O_0[t]$
end for
Verify atomic encoding in complete transcriptions
return $\tau_1, \mathcal{V}_1, E_1, O_1$

This changes each target character from $k(c)$ within-character autoregressive decisions to a single atomic prediction,

$$\ell_{\text{atom}}(c \mid x_c) = -\log p_{\Theta_1}(u_c \mid x_c), \tag{5}$$

while full-page generation remains autoregressive.

Constituent-anchored initialization. Each new atomic unit is initialized from its original constituent tokens:

$$E_1^{(0)}[u_c] = \frac{1}{k(c)} \sum_{j=1}^{k(c)} E_0[t_j^c],$$
$$O_1^{(0)}[u_c] = \frac{1}{k(c)} \sum_{j=1}^{k(c)} O_0[t_j^c]. \tag{6}$$

Qwen3.5-4B uses tied input/output representations before adaptation, and vocabulary resizing preserves this tying, so the two initializations coincide at construction time. The initialization is specific to the atomic target rows; it does not imply that only those rows are trainable.

Language-side adaptation. Introducing atomic output units requires extending the token-indexed input/output representations, so the associated parameter growth is intrinsic to the output-space intervention. We freeze the visual encoder and multimodal projector, apply LoRA (Hu et al., 2022) to the language backbone, and fully train `embed_tokens` and `lm_head` during Stage 2. This removes the initial weight tying, allowing E_1 and O_1 to update independently across all rows. Detailed parameter accounting and training settings are provided in Appendix D.

3.3 Contextualized Page-Level Supervision

Atomic output units determine how a character is represented, but not where it belongs in a page transcription. Historical Chinese OCR additionally requires reading order and structural relations among vertical columns, main text, and annotations. We therefore learn GAOSA in complete page contexts rather than isolated glyphs.

Layout-conditioned synthesis. We construct synthetic pages from a structured specification

$$Z = \{(c_i, \mathbf{b}_i, \rho_i)\}_{i=1}^{N}, \tag{7}$$

where c_i is a character, $\mathbf{b}_i$ its region, and ρ_i its structural role. An LLM-based planner proposes page layouts, which are populated with historical fonts and appearance degradations. The image and canonical transcription are generated from the same specification, providing aligned supervision for layout and reading order; layout metadata are used only for data construction and are never exposed to the OCR model.

Effective target-character exposure. Atomic units must occur at supervised response positions to receive useful gradients. After preprocessing and truncation, let $\mathbf{t} = \tau_1(y)$ and $\mathcal{M}_y$ denote retained response-token positions. We define

$$n_{\mathcal{D}}^{\mathrm{sup}}(c) = \sum_{(I,q,y)\in\mathcal{D}} \sum_{j\in\mathcal{M}_y} \mathbf{1}[t_j = u_c]. \tag{8}$$

This statistic guides data construction rather than introducing an auxiliary loss: underexposed characters are preferentially synthesized so atomic units receive direct image–text supervision in page contexts. Synthetic-data, annotation, and target-character construction details are provided in Appendix A.

3.4 Two-Stage Page-to-Sequence Learning

Stage 1 retains τ_0 and performs SFT on synthetic and manually annotated historical pages, yielding the page-adapted checkpoint Θ_{page}. Stage 2 starts from this checkpoint, constructs τ_1, initializes atomic units with Equation 6, and trains on target-character-enriched synthetic and annotated real pages, denoted by fragmentation-enriched supervision $\mathcal{S}_f$. The provenance of the manually annotated collection is given in Appendix B.

For the principal Stage 2 comparison, conventional SFT and GAOSA start from the same Θ_{page} and receive the same Stage 2 examples and supervision. Conventional SFT retains fragmented paths under τ_0, whereas GAOSA uses atomic units under τ_1 together with the language-side adaptation above. Because GAOSA also expands, unties, and fully optimizes the embedding/output matrices, this is a *data-matched* comparison rather than a parameter- or optimization-matched isolation of tokenization alone.

Let Φ_{page} denote Stage 1 parameters frozen during Stage 2 and Ω the trainable language-side parameters. For $\mathbf{t} = \tau_1(y)$, Stage 2 minimizes

$$\mathcal{L}_{\mathrm{page}}(\Omega) = -\frac{1}{N_{\mathrm{sup}}} \sum_{(I,q,y)\in\mathcal{D}_2} \sum_{j\in\mathcal{M}_y} \log p_{\Phi_{\mathrm{page}},\Omega}(t_j \mid I, q, \mathbf{t}_{<j}). \tag{9}$$

Here N_{sup} counts supervised response-token positions; instruction, image-placeholder, and padding positions are excluded, and training uses ground-truth response prefixes. Full output-space and training configurations are provided in Appendix D.

4 Evaluation Datasets

We evaluate four historical Chinese benchmarks. For full-page evaluation on public document images, we use 2,198 pages from AncientDoc (Yu et al., 2025) and 2,198 pages from MTHv2 (Ma et al., 2020). Their original annotations are not consistently suitable for our page-to-sequence protocol because they do not provide a complete serialized transcription of all page content, including headers, footers, and other page elements. We therefore manually re-annotate complete page-level transcription sequences for the selected images and evaluate all systems against the same references; the annotation protocol is illustrated in Figure 6 of Appendix B.1. These transcriptions will be released. HCLDoc-582 contains 582 pages with 291 annotated and 291 synthetic. FELDoc-1K contains 1,000 synthetic pages enriched with fragmented targets and supports both target-character and page-level evaluation.

For cross-script evaluation, HHD-Ethiopic (Belay et al., 2024) provides 57,374 training examples and a 15,935-example `test_18th` split. OCR-IPA (Okabe et al., 2026) uses 15,000 training and 1,000 evaluation examples from the `synth` collection. Evaluation pairs are held out from training.

Table 1: Benchmarks used in this work. *Target* denotes target-character evaluation. † Public document images evaluated with our manually re-annotated complete page-level transcriptions.

Dataset	Samples	Type	Focus
AncientDoc†	2,198	Public	Page
MTHv2†	2,198		Page
HHD-Ethiopic	15,935		Target
OCR-IPA	1,000		Target
FELDoc-1K	1,000	In-house	Target+Page
HCLDoc-582	582		Page

Metrics are computed on decoded character sequences rather than token IDs. For fragmented-character evaluation, precision (P), recall (R), and F1 are computed independently for each evaluation sample and then averaged across samples; all three metrics are reported as percentages. Consequently, the reported F1 need not equal the harmonic mean of the reported averaged precision and recall. For full-page transcription, NED, AR, and CR are likewise computed per page and averaged across pages. AR and CR are reported as percentages, whereas NED is reported as a unitless decimal rounded to two decimal places. Fragmentation groups are fixed by encoding length under τ_0. In the historical Chinese experiments, *Stage-1* denotes the page-adapted Qwen3.5-4B checkpoint, τ_1 the glyph-aligned output space, and $\mathcal{S}_f$ the fragmentation-enriched Stage 2 supervision.

5 Experiments

5.1 Experimental Setup

We follow the datasets, metrics, and notation in Section 4. Unless stated otherwise, k is computed with the pre-adaptation tokenizer τ_0, and the resulting character groups remain fixed across variants. P, R, and F1 diagnose recovery of the designated target-character inventory, whereas NED, AR, and CR evaluate complete-page transcription.

We compare against representative general-purpose VLMs, including GPT-5.6 Sol (OpenAI, 2026), Qwen3.6-27B (Qwen Team, 2026b), Qwen3.7-Plus (Qwen Team, 2026c), and Qwen3.8-Max (Qwen Team, 2026d); OCR-specialized systems, including DeepSeek-OCR-2 (Wei et al., 2026), dots.ocr (Li et al., 2025), MinerU2.5-2509-1.2B (Niu et al., 2025), and PaddleOCR-VL-1.6 (Zhang et al., 2026); and TongGu-VL-2B (Cao et al., 2025), which is specialized for Chinese classical documents. General-purpose VLMs use the same OCR instruction as our model, whereas OCR-specialized systems use their official prompts and inference interfaces because they are trained around their native OCR protocols. PaddleOCR-VL-1.6 is evaluated with its complete official pipeline rather than only the standalone VLM component. All outputs are scored in the same decoded-character evaluation space. For backbone validation, we additionally instantiate GAOSA

Table 2: **Fragmented-character recognition on FELDoc-1K.** All values are percentages. P, R, and F1 are sample-wise averaged over target characters with $k > 1$. F1@k reports F1 for characters represented by exactly k tokens under τ_0; F1@1 is shown in gray as a non-fragmented reference.

Model	Fragmented Characters ($k > 1$)			F1 by Fragmentation Degree			
	P ↑	R ↑	F1 ↑	F1@1	F1@2	F1@3	F1@4
General VLMs							
GPT-5.6 Sol	12.29	2.23	3.74	86.47	9.31	0.10	0.00
Qwen3.6-27B	21.21	13.62	16.56	94.24	40.41	2.53	0.01
Qwen3.7-Plus	26.95	14.87	19.12	94.48	46.07	4.13	0.24
Qwen3.8-Max	26.90	17.57	21.22	94.41	51.31	4.56	0.03
OCR Specialists							
DeepSeek-OCR-2	12.28	0.95	1.70	53.44	4.22	0.03	0.00
dots.ocr	13.54	6.42	8.68	90.69	21.70	0.25	0.00
MinerU2.5-2509-1.2B	10.06	2.14	3.47	78.47	8.64	0.06	0.00
PaddleOCR-VL-1.6	30.77	18.38	22.95	93.01	56.18	2.49	0.00
Historical OCR / Adapted Backbones							
TongGu-VL-2B	10.61	3.93	5.70	82.73	14.24	0.05	0.02
Qwen3.5-4B (Stage-1)	10.25	6.11	7.64	93.43	18.89	0.81	0.00
Qwen3.5-4B w/ GAOSA	91.00	86.28	88.55	99.02	93.23	**95.99**	78.40
InternVL3.5-4B (Base)	34.67	19.97	25.21	92.34	29.79	22.04	0.14
InternVL3.5-4B w/ GAOSA	**94.55**	**92.96**	**93.74**	**99.53**	**95.35**	90.97	**92.19**

on InternVL3.5-4B (Wang et al., 2025); the controlled evidence for GAOSA itself is provided by the data-matched Stage 2 comparison described in Section 5.3.

5.2 Fragmented-Character Recognition

5.2.1 Diagnostic Results on FELDoc-1K

Table 2 reports GAOSA results on both Qwen3.5-4B and InternVL3.5-4B, which serve different experimental roles. Qwen3.5-4B is the backbone of our complete historical OCR system; it therefore first undergoes large-scale Stage 1 page-to-sequence training to acquire full-page layout understanding and transcription capability, followed by GAOSA adaptation in Stage 2. The resulting model is further evaluated as our complete method against other historical document parsing systems in Section 5.5. InternVL3.5-4B provides a complementary backbone validation: we do not apply the Stage 1 historical-page adaptation, but instead introduce GAOSA directly from the Base model to test whether the same output-space adaptation transfers to a different VLM backbone.

Across external models, we observe a highly consistent fragmentation-dependent degradation. These models generally recognize non-fragmented characters well, but performance drops sharply as the number of tokens required to represent a single character increases. For example, Qwen3.8-Max decreases from 94.41% F1 at $k = 1$ to 4.56% at $k = 3$ and 0.03% at $k = 4$, while PaddleOCR-VL-1.6 similarly drops from 93.01% to 2.49% and 0.00%. The same trend appears across other general-purpose VLMs, OCR-specialized systems, and the historical OCR model. This suggests that these models are not uniformly weak at character recognition; rather, recognition becomes substantially more difficult when a visually atomic character must be generated through an increasingly fragmented output sequence.

GAOSA substantially changes this pattern. For Qwen3.5-4B, Stage 1 already achieves 93.43% F1 on non-fragmented characters, yet F1@2, F1@3, and F1@4 remain only 18.89%, 0.81%, and 0.00%, respectively, indicating that page-to-sequence alignment alone does not resolve output fragmentation. After introducing GAOSA, these scores increase to 93.23%, 95.99%, and 78.40%, with an overall F1 of 88.55% on fragmented characters. On InternVL3.5-4B Base, GAOSA similarly raises F1@2/3/4 from 29.79%/22.04%/0.14% to 95.35%/90.97%/92.19%, yielding an overall fragmented-character F1 of 93.74%. These results show that GAOSA effectively mitigates the recognition degradation associated with increasing token fragmentation, both within our full two-stage historical OCR pipeline and when transferred directly to a different VLM backbone.

Table 3: **Data-matched Stage-2 comparison on FELDoc-1K ($k > 1$).** All rows start from the same Qwen3.5-4B Stage-1 checkpoint; SFT and GAOSA use identical Stage-2 training examples.

Method	τ_1	$\mathcal{S}_f$	P ↑	R ↑	F1 ↑
Stage-1	✗	✗	10.25	6.11	7.64
$+\tau_1$	✓	✗	10.20	6.09	7.61
$+\mathcal{S}_f$ (SFT)	✗	✓	62.80	56.91	59.68
GAOSA	✓	✓	**91.00**	**86.28**	**88.55**

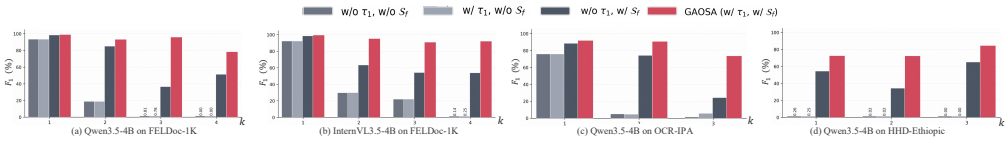

Figure 2: **Fragmentation-stratified F1 across backbones and scripts.** F1@k groups characters by their token count k under the pre-adaptation tokenizer τ_0. (a) FELDoc-1K with Qwen3.5-4B: data-matched Stage 2 SFT versus GAOSA. (b) FELDoc-1K with InternVL3.5-4B: Base versus GAOSA, used as a backbone-compatibility test. (c–d) OCR-IPA and HHD-Ethiopic with Qwen3.5-4B: data-matched SFT versus GAOSA.

5.3 Data-Matched Stage-2 Comparison

The central comparison mirrors the two-stage design in Section 3.4. All rows start from the same Qwen3.5-4B Stage-1 checkpoint, which provides a shared page-aligned initialization before Stage 2 fragmented-character adaptation. The $+\mathcal{S}_f$ condition performs conventional SFT under τ_0, while GAOSA uses the same Stage 2 data and supervision under τ_1. Thus, the comparison is matched in initialization and supervision, with GAOSA differing in its atomic output-space parameterization and associated output-layer optimization.

Table 3 contrasts targeted Stage 2 supervision with the complete GAOSA intervention. Introducing τ_1 without fragmentation-enriched supervision leaves F1 unchanged (7.64 vs. 7.61), showing that atomic output construction and initialization alone are insufficient. Conventional Stage 2 SFT with $\mathcal{S}_f$ raises F1 to 59.68%, while GAOSA reaches 88.55% on the same data, a further 28.87-point gain.

These results show that effective Stage 2 supervision is required to learn the new representations and that the complete GAOSA intervention substantially outperforms conventional SFT under data-matched supervision. We next test whether the gains concentrate on fragmented groups.

Figure 1 examines whether the gains from GAOSA are concentrated on tokenizer-fragmented characters rather than being uniformly distributed across all characters. For Qwen3.5-4B, the data-matched Stage 2 SFT baseline ($+\mathcal{S}_f$) already reaches 98.46% at $k = 1$, and GAOSA increases it only slightly to 99.02%. In contrast, F1@3 rises from 36.54% to 95.99%, while F1@4 improves from 51.37% to 78.40%. InternVL3.5-4B exhibits the same trend, with F1@2/3/4 increasing from 29.79/22.04/0.14% to 95.35/90.97/92.19%, respectively. We observe the same pattern across different writing systems: on HHD-Ethiopic and OCR-IPA, GAOSA likewise provides larger gains for character groups with higher degrees of tokenizer fragmentation. Because the InternVL experiment starts directly from the Base model rather than from a matched SFT baseline, it mainly serves to evaluate the transferability of GAOSA across different backbones. Performance does not necessarily vary monotonically with k, since character composition and training exposure also differ across fragmentation groups.

5.4 Cross-Script Validation

We next test whether the same intervention transfers beyond historical Chinese. These are not zero-shot experiments: on each dataset, GAOSA is compared with a data-matched conventional SFT baseline receiving the same supervision but retaining the original output space.

As shown in Table 4, GAOSA improves F1 from 73.93% to 90.78% on OCR-IPA and from 42.56% to 76.17% on HHD-Ethiopic. Precision and recall increase together on both datasets, so the gains

Table 4: **Cross-script validation on fragmented target characters ($k > 1$).** Within each dataset, SFT and GAOSA use identical training examples and supervision. P, R, and F1 are sample-wise averaged.

Dataset	Model	P ↑	R ↑	F1 ↑
OCR-IPA	Qwen3.5-4B (SFT)	75.94	73.94	73.93
	w/ GAOSA	**90.28**	**92.19**	**90.78**
HHD-Ethiopic	Qwen3.5-4B (SFT)	44.44	42.28	42.56
	w/ GAOSA	**78.02**	**75.76**	**76.17**

Table 5: **Full-page transcription across four historical Chinese benchmarks.** AR and CR are reported as percentages and are higher-is-better; NED is reported as a unitless decimal and is lower-is-better. Each metric is computed independently on each complete decoded page transcription and then averaged across pages.

Model	HCLDoc-582			AncientDoc			MTHv2			FELDoc-1K		
	AR ↑	CR ↑	NED ↓	AR ↑	CR ↑	NED ↓	AR ↑	CR ↑	NED ↓	AR ↑	CR ↑	NED ↓
General VLMs												
Qwen3.6-27B	68.49	77.46	0.27	79.65	87.26	0.18	78.85	84.43	0.21	16.00	16.05	0.19
Qwen3.7-Plus	75.75	80.73	0.24	**83.43**	88.86	0.16	79.08	85.22	0.21	16.91	16.91	0.19
Qwen3.8-Max	77.75	**88.66**	0.20	61.12	89.28	0.16	37.24	**86.60**	0.19	17.20	17.34	0.19
OCR Specialists												
DeepSeek-OCR-2	-8.67	10.22	0.92	-1.17	12.02	0.90	-7.38	20.72	0.84	0.41	0.50	0.92
dots.ocr	17.74	69.71	0.40	33.07	86.71	0.24	4.59	80.31	0.33	6.19	6.25	0.22
MinerU2.5-2509-1.2B	-120.75	35.60	0.74	-48.53	51.93	0.63	19.88	56.76	0.56	1.42	1.42	0.72
PaddleOCR-VL-1.6	55.72	72.51	0.39	77.77	88.74	0.19	46.13	73.58	0.38	16.81	16.92	0.19
Historical OCR Models												
TongGu-VL-2B	-18.60	39.89	0.64	53.27	66.38	0.37	-49.79	47.30	0.65	2.02	3.17	0.40
Qwen3.5-4B w/ GAOSA	**85.76**	87.31	**0.14**	76.27	**91.43**	**0.11**	**83.38**	85.38	**0.16**	**88.69**	**88.76**	**0.03**

are not explained by a precision–recall trade-off. Together with the second-backbone result, these experiments show that glyph-aligned output adaptation is not specific to historical Chinese.

5.5 Full-Page Historical Document Transcription

Finally, we test whether the character-level intervention remains compatible with page-to-sequence transcription. Table 5 compares the final Qwen3.5-4B with GAOSA against external systems on complete historical pages. GAOSA obtains the lowest NED among the evaluated systems on all four benchmarks: 0.14 on HCLDoc-582, 0.11 on AncientDoc, 0.16 on MTHv2, and 0.03 on FELDoc-1K. Its AR/CR rankings are strong but not uniformly best: Qwen3.7-Plus has higher AR on AncientDoc, while Qwen3.8-Max has higher CR on HCLDoc-582 and MTHv2.

Negative AR can occur because AR is not clipped at zero; severe over-generation can make edit distance exceed the reference length. Overall, the results show that GAOSA's fragmented-character gains coexist with strong complete-page transcription, supporting the two-stage design rather than reducing the method to isolated-character recognition.

6 Conclusion

We identify token fragmentation as an output-representation bottleneck in generative OCR: a visually atomic glyph may require multiple autoregressive predictions, leaving substantial residual errors even after supervised fine-tuning. We introduce GAOSA, which replaces fragmented character paths with atomic output units while preserving page-to-sequence generation. Under data-matched Stage 2 supervision on FELDoc-1K, conventional SFT reaches 59.68% fragmented-character F1, while GAOSA reaches 88.55%. The gains transfer to IPA and Ethiopic, remain compatible with a second VLM backbone, and coexist with strong full-page transcription. These results support aligning output granularity with perceptual units as an effective direction for generative OCR.

Reproducibility Statement

We provide detailed descriptions of the data-construction pipeline, annotation protocol, data sources, atomic output-space construction, and two-stage training configuration in Appendices A, B, and D. The appendices report the training-data composition, tokenizer and output-space construction, trainable modules, optimization settings, sequence lengths, and hardware configuration. Evaluation datasets, metric definitions, aggregation protocols, and baseline inference settings are specified in Sections 4 and 5.1. The re-annotated page-level transcriptions and benchmarks described in this work will be released to support reproducibility.

AI Assistance

Generative AI tools were used to assist with language editing and polishing of author-written text. We also used an LLM to generate structured layout templates as part of the synthetic-data construction pipeline described in Appendix A. All AI-assisted outputs used in the research or manuscript were reviewed by the authors. The authors take full responsibility for the methodology, experimental results, claims, and final content of this work.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Birhanu Hailu Belay, Isabelle Guyon, Tadele Mengiste, Bezawork Tilahun, Marcus Liwicki, Tesfa Tegegne, and Romain Egele. A historical handwritten dataset for ethiopic ocr with baseline models and human-level performance. In *Document Analysis and Recognition – ICDAR 2024*, pp. 23–38. Springer, 2024. doi: 10.1007/978-3-031-70543-4_2.

Jiahuan Cao, Yang Liu, Peirong Zhang, Yongxin Shi, Kai Ding, and Lianwen Jin. TongGu-VL: Advancing visual-language understanding in chinese classical studies through parameter sensitivity-guided instruction tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11111–11120, 2025.

Konstantin Dobler and Gerard de Melo. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13440–13454, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.829. URL https://aclanthology.org/2023.emnlp-main.829/.

Haisu Guan, Jinpeng Wan, Yuliang Liu, Pengjie Wang, Kaile Zhang, Zhebin Kuang, Xinyu Wang, Xiang Bai, and Lianwen Jin. An open dataset for the evolution of oracle bone characters: EVOBC. *arXiv preprint arXiv:2401.12467*, 2024. doi: 10.48550/arXiv.2401.12467. URL https://arxiv.org/abs/2401.12467v1.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. Ocr improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1164–1174, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.92.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *Computer Vision – ECCV 2022*, volume 13688 of *Lecture Notes in Computer Science*, pp. 498–517, Cham, 2022a. Springer. doi: 10.1007/978-3-031-19815-1_29. URL https://doi.org/10.1007/978-3-031-19815-1_29.

Geonuk Kim, Jaemin Son, Kanghyu Lee, and Jaesik Min. Character decomposition to resolve class imbalance problem in Hangul OCR. *arXiv preprint arXiv:2208.06079*, 2022b. doi: 10.48550/arXiv.2208.06079. URL https://arxiv.org/abs/2208.06079.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. Efficient and effective vocabulary expansion towards multilingual large language models. *arXiv preprint arXiv:2402.14714*, 2024. doi: 10.48550/arXiv.2402.14714. URL https://arxiv.org/abs/2402.14714.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13094–13102, 2023. doi: 10.1609/aaai.v37i11.26538. URL https://ojs.aaai.org/index.php/AAAI/article/view/26538.

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*, 2025. URL https://arxiv.org/abs/2512.02498.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. Joint layout analysis, character detection and recognition for historical document digitization. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 31–36, 2020. doi: 10.1109/ICFHR2020.2020.00017.

Benjamin Minixhofer, Edoardo M. Ponti, and Ivan Vulić. Zero-shot tokenizer transfer. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1484. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/532ce4fcf853023c4cf2ac38cbc5d002-Abstract-Conference.html.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, et al. MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*, 2025. URL https://arxiv.org/abs/2509.22186.

Shu Okabe, Dejvi Zelo, and Alexander Fraser. Optical character recognition for the international phonetic alphabet. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 265–273. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-short.19.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026. URL https://openai.com/index/gpt-5-6/.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1608. URL https://arxiv.org/abs/2305.15425.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.7-Plus. Alibaba Cloud Model Studio documentation, 2026c. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-7-plus. Accessed 2026-09-26.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026d. URL https://qwen.ai/blog?id=qwen3.8.

Antonio Ríos-Vila, Jorge Calvo-Zaragoza, and Thierry Paquet. Sheet music transformer: End-to-end optical music recognition beyond monophonic transcription. In *Document Analysis and Recognition – ICDAR 2024*, volume 14809 of *Lecture Notes in Computer Science*, pp. 20–37, Cham, 2024. Springer. doi: 10.1007/978-3-031-70552-6_2. URL https://doi.org/10.1007/978-3-031-70552-6_2.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024. doi: 10.48550/arXiv.2409.01704. URL https://arxiv.org/abs/2409.01704.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR 2: Visual causal flow. *arXiv preprint arXiv:2601.20552*, 2026. URL https://arxiv.org/abs/2601.20552.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. How can we effectively expand the vocabulary of LLMs with 0.01GB of target language text? *Computational Linguistics*, 52(1):295–330, March 2026. doi: 10.1162/COLI.a.581. URL https://doi.org/10.1162/COLI.a.581.

Haiyang Yu, Yuchuan Wu, Fan Shi, Lei Liao, Jinghui Lu, Xiaodong Ge, Han Wang, Minghan Zhuo, Xuecheng Wu, Xiang Fei, Guozhi Tang, An-Lan Wang, Hanshen Zhu, Yangfan He, Quanhuan Liang, Liyuan Meng, Chao Feng, Can Huang, Jingqun Tang, and Bin Li. Benchmarking vision-language models on chinese ancient documents: From ocr to knowledge reasoning. *arXiv preprint arXiv:2509.09731*, 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. doi: 10.48550/arXiv.2606.03264. URL https://arxiv.org/abs/2606.03264.

A Data Construction

A.1 Synthetic Data Pipeline

Full-page transcription requires page-level supervision at a scale that manual annotation alone cannot provide. We therefore synthesize historical pages from structured layout specifications. Figure 3 illustrates the generation pipeline, and Figure 4 shows representative generated pages.

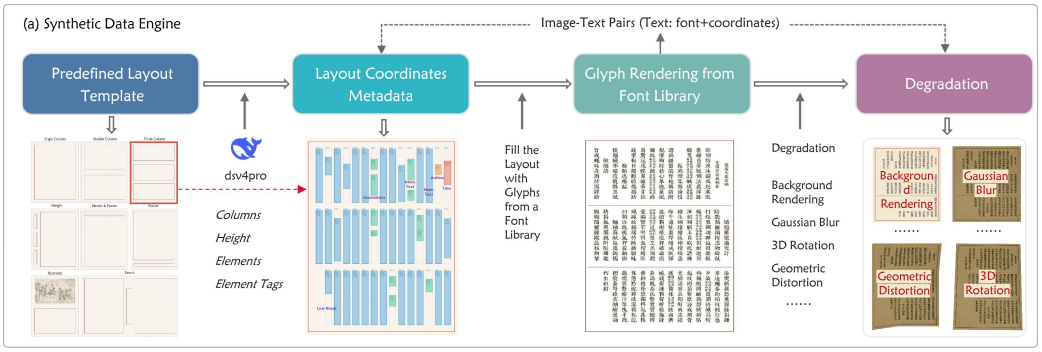

Figure 3: Synthetic historical page generation pipeline. The deterministic rendering process preserves an exact correspondence between each generated page image and its canonical transcription, providing scalable image-text supervision for historical OCR.

Layout-template generation. Layout specifications are generated as JSON templates using DeepSeek-V4-Pro. The `Tag`, `counts`, and `begins` fields of each specification are further edited to obtain renderable layouts. The resulting specifications span fourteen layout families, covering single-, two-, and three-column page structures, book spreads, annotation-dense layouts, poetry blocks with decorative patterns, and various punctuation placements.

Text insertion and rendering. Given the generated layout specifications and transcription sequences, texts are inserted into the corresponding regions according to the layout constraints. The pages are rendered with historical Kai-style fonts, including an Extension-B font to ensure that rare glyphs are rendered faithfully instead of being replaced by fallback symbols. The resulting collection contains 34,796 base page images.

Degradation augmentation. To reduce the gap between synthetic renderings and real historical scans, we apply visual degradation to the rendered pages. The degradation process simulates common artifacts in historical documents, including paper aging, blur, ink variation, and structural damage. These transformations modify only the visual appearance while preserving the canonical transcription.

The final synthesis pipeline produces 74,520 synthetic historical pages for page-level supervision.

A.2 Manual Annotation

The annotated collection contains 18,432 page images with reference transcriptions prepared by five annotators following the canonical transcription protocol. The annotation process involved approximately 3,168 person-hours over three consecutive calendar months.

Annotators were trained to follow the unified transcription protocol, including historical reading order, annotation placement, line-break conventions, and removal of non-transcriptional elements such as headers, footers, page numbers, and the central book-fold area. The annotation process was supported by automatic quality checking and manual sampled review to identify inconsistencies and ensure transcription consistency.

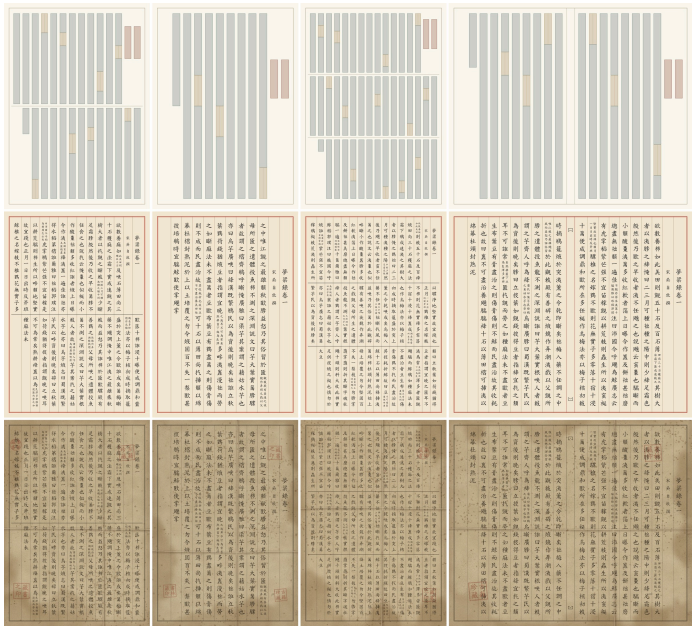

Figure 4: Representative examples from the synthetic page generation pipeline. From top to bottom: JSON-defined layout templates, clean renderings, and the corresponding degraded pages used for training.

A.3 Target-Character-Enriched Synthetic Data

The target inventory contains 34,013 historical character forms. Characters are grouped according to their standalone encoding length under the original tokenizer. Table 6 summarizes the inventory composition and targeted synthetic-page allocation.

Table 6: Target-character inventory and second-stage synthetic-page allocation. Fragmentation degree is the number of original tokens per character.

Original tokens per character	Target characters	Targeted synthetic pages
2	14,187	5,000
3	5,079	6,483
4	14,747	6,500
Total	**34,013**	**17,983**

The target-character-enriched corpus differs from the general synthetic collection used for page-level alignment. Instead of providing only layout diversity, these pages explicitly introduce fragmented target characters into the rendered texts, providing contextual supervision for vocabulary adaptation.

The corpus contains 17,983 targeted synthetic pages. Together with the 18,432 manually annotated pages, the second-stage adaptation corpus contains 36,415 image-text pairs.

B Data Sources

B.1 Annotated Historical Collection

The manually annotated collection contains 18,432 historical page images with canonical transcriptions following the unified transcription protocol. The collection is constructed from two historical document sources: the Shuge collection and the National Library of China (NLC). The annotation guidelines are illustrated in Figure 6.

Table 7: Source composition of the manually annotated historical collection.

Source	Description	Pages
Shuge digital library	59 title or edition entries	10,621
National Library of China	Three batches of holdings	7,811
Total		**18,432**



Figure 5: Annotation process for the manually transcribed historical collection.

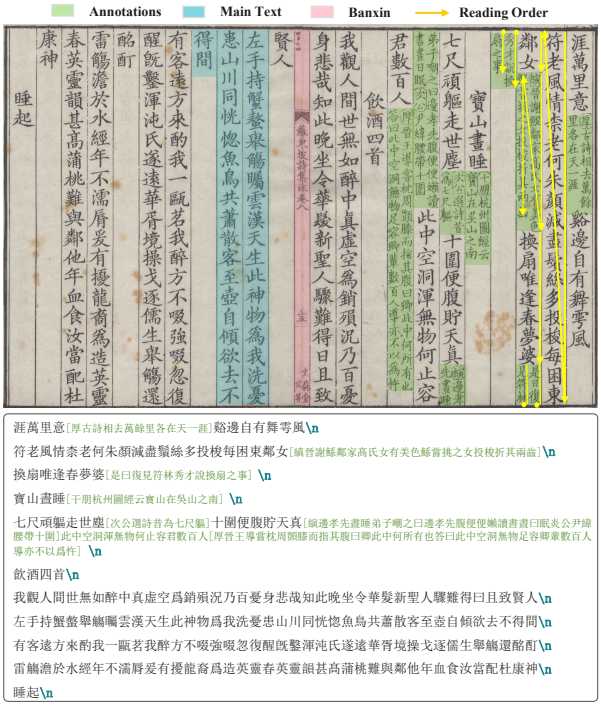

Figure 6: Page-to-sequence transcription for Chinese historical OCR. Non-transcriptional elements such as banxin, headers, and footers are removed as noise.

The Shuge portion covers diverse historical documents, including classics, histories, medical and agricultural treatises, literary collections, and bibliographic works. The National Library of China portion consists of three batches of holdings covering representative historical collections.



Figure 7: Model comparison on a real historical page from MTHv2. The example contains vertically arranged text columns, highlighting differences in reading-order recovery and final transcription quality.

C Qualitative Results

C.1 Real Historical Pages (MTHv2)

Figure 7 compares model outputs on a real historical page from MTHv2. The page contains vertically arranged text columns, requiring the model to recover historical reading order and transform the two-dimensional page structure into a coherent transcription sequence. The example illustrates the page-to-sequence capability learned through image-text supervision: the proposed model preserves the column order and produces a transcription closely aligned with the reference, while baseline systems exhibit character substitution, omission, or reading-order errors.

C.2 Fragmented-Character Pages (FELDoc-1K)

Figure 9 presents a representative FELDoc-1K page enriched with fragmented characters under the original tokenizer. The highlighted regions indicate target characters affected by token fragmentation. Conventional SFT models often substitute or omit these characters, whereas GAOSA recovers a larger fraction of fragmented targets by introducing atomic output representations.

Input ʁɪçtən-fɛtlaɪ̯bɪk-fɛɐ̯tik-knaɪpmʃtɛʁbm, tiːfkyːlpɪʦ͡ʃaˈbəkantn̩ ʔaʊ̯sʃliːse

OCR outputs

GT	ʁɪçtən-fɛtlaɪ̯bɪk-fɛɐ̯tik-knaɪpmʃtɛʁbm, tiːfkyːlpɪʦ͡ʃaˈ bəkantn̩ ʔaʊ̯sʃliːse
GAOSA	ʁɪçtən-fɛtlɑɪ̯bɪk-fɛɐ̯tik-knaɪpmɻtɛʁbm, tiːfkyːlpɪʦ͡ʃaˈ bəkantn̩ ʔaʊ̯sʃliːse
SFT	ʁiçtən-fɛtlɑɪ̯k-fɛɐ̯tik-knaɪpɻmʃtɛbɒm, tiːfkyːlpɪɾsaˈ bəkantŋ ʔaʊsʌiːse
Exp. only	biçtən-fetlaibik-feətik-knaipmftexbm, tiːfkyːlpiʦ͡ʃaˈbəkantnʔauɣsʃliːse
Base	biçtən-fetlaibik-feətik-knaipmftexbm, tiːfkyːlpiʦ͡ʃaˈbəkantnʔauɣsʃliːse

Input ካርዩ·ክሞሁ፡በውስ ወጠጽኡ፡ፇኣኝሳዩ፡ኣ

OCR outputs

GT	ከ ር ዪ ፡ ስ ም ሁ ፡ በ ው ስ ፡ ፡ ወ መ ጽ ኡ ፡ ፪ ኣ እ ማ ድ ፡ እ
GAOSA	ከ ር ዪ ፡ ስ ም ሁ ፡ በ ው ስ ፡ ወ መ ጽ ኡ ፡ ፪ ኣ እ ግ ድ ፡ እ
SFT	ሰ ር ቆ ፡ ከ ም 0 ፡ በ ው ስ ጢ፡ ወ መ ን እ ሲ ፡ ፮ ፱ እ ግ ዚ ኣ ብ ሐ
Exp. only	In CF·h q° v.: ΠΦ·h 0 <Φ πλ·: ελχσς
Base	In Ck·h P° v.: ΠΦ·h 0 <0 πλ·: ελχσς

Figure 8: Cross-script qualitative examples of output-space extension on OCR-IPA and HHD-Ethiopic. Characters requiring two or more tokens under the original tokenizer are highlighted in red in the ground-truth transcription. In the model outputs, correctly recognized target characters are highlighted in green. The examples illustrate the effect of output-space adaptation on fragmented visual units across different writing systems.



Figure 9: Qualitative comparison on a representative FELDoc-1K page containing fragmented characters. Red highlights indicate target characters or incorrect predictions, while green highlights indicate correctly recovered targets. The example illustrates the benefit of converting fragmented glyphs into atomic output tokens.

C.3 Cross-Script Atomic Output-Space Construction

Figure 8 presents qualitative examples from OCR-IPA and HHD-Ethiopic to examine whether the tokenization bottleneck observed in Chinese historical OCR extends to other writing systems. Although these scripts have different character inventories and visual structures, visually atomic units may still be fragmented into multiple output tokens under the original tokenizer. The examples show that output-space extension can alleviate this representation mismatch beyond Chinese historical characters.

D Experimental Details

D.1 Atomic Output-Space Construction

We start from a historical character inventory and analyze its tokenization under the original tokenizer. Characters whose representations require multiple tokens are selected for atomic output-space adaptation. We identify 34,013 fragmented characters, including 14,187 characters represented by two tokens, 5,079 by three tokens, and 14,747 by four tokens.

The original tokenizer contains 248,044 BPE entries and 33 special tokens (maximum token ID 248,076). We append the 34,013 target characters as ordinary tokens with consecutive IDs from 248,077 to 282,089. Existing token IDs and special-token assignments remain unchanged. The embedding matrix is resized from 248,320 to 282,090 rows. Due to the existing padding rows in the original embedding matrix, 33,770 rows are physically appended.

To provide sufficient supervision for the atomic output space, we construct an additional target-character-enriched adaptation corpus using the same synthesis pipeline. Unlike the general synthetic collection used for page-level alignment, these pages are newly generated by explicitly injecting target fragmented characters into rendered texts. The corpus contains 17,983 synthetic pages. Together with 18,432 manually annotated pages, the adaptation corpus contains 36,415 image-text pairs.

D.2 Training Configuration

The model is trained with a two-stage alignment framework. In Stage 1, we perform Page-to-Sequence Alignment starting from the pretrained Qwen3.5-4B model. The model is trained with large-scale synthetic historical pages and 18,432 manually annotated pages to learn the mapping from full-page images to canonical transcription sequences, obtaining a page-aligned checkpoint.

In Stage 2, we start from this checkpoint, extend the token-indexed output space with dedicated historical-character tokens, and perform Glyph-to-Token Alignment using a target-character-enriched adaptation corpus. The Stage-2 corpus contains 36,415 image-text pairs, including 18,432 manually annotated pages and 17,983 newly generated synthetic pages.

We use Qwen3.5-4B as the backbone of the two-stage framework. During both stages, the vision tower and multimodal projector are frozen. LoRA is applied to all language-model linear modules. During Stage 2, the token embeddings and language-model head are additionally optimized together with LoRA to adapt the expanded vocabulary.

Table 8: Training configurations for the two-stage alignment framework.

Component	**Stage 1**	**Stage 2**
Initialization	Qwen3.5-4B	Page-aligned checkpoint
Data	74,520 synthetic + 18,432 manual pages	17,983 targeted synthetic + 18,432 manual pages
Framework	LLaMA-Factory	LLaMA-Factory
Data format	ShareGPT	ShareGPT
Template	`qwen3_5_nothink`	`qwen3_5_nothink`
LoRA rank	8	8
LoRA alpha	16	16
LoRA target	all linear modules	all linear modules
Trainable modules	LoRA	LoRA + embed tokens + LM head
Optimizer	Adafactor	Adafactor
Learning rate	5×10^{-5}	5×10^{-5}
Scheduler	cosine decay	cosine decay
Warmup ratio	0.03	0.03
LoRA dropout	0	0
Max gradient norm	1.0	1.0
Precision	pure bf16	pure bf16
Sequence length	2048 tokens	1792 tokens
Gradient checkpointing	enabled	enabled
Hardware	Ascend 910B4 $\times$ 4	Ascend 910B4 $\times$ 6