AnandaSky: A Vision–Language Model for Line-Level Transcription of Historical Sinographic Documents

Colin Brisson[1,2], **Ayoub Kahfy**[2], **Frédéric Constant**[1,3], **Marc Bui**[4]
[1]CRCAO, EPHE–PSL; [2]Badiane; [3]ERMES, Université Côte d'Azur; [4]AOROC, EPHE–PSL
{colin, ayoub}@badiane.ai, frederic.constant@univ-cotedazur.fr, marc.bui@ephe.psl.eu

Abstract

We present AnandaSky, a vision–language model for line-level transcription of historical sinographic documents. The model combines a compact high-resolution visual encoder with global attention, 10 px patches, uncompressed visual prefix and a Qwen3-0.6B autoregressive decoder. It is trained at scale on 4M annotated lines from documents produced in China and Korea between the 8th and 20th centuries. Across in-domain and held-out public benchmarks, AnandaSky achieves sub-1% CER on five of eight datasets, sets a new state of the art on MTHv2 with 0.92% CER, and shows strong transfer to unseen collections. For EvaHan 2026, full fine-tuning on the organizers' data to match task-specific annotation conventions reduces CER relative to the official baseline by 5.2% on prints and 12.1% on manuscripts, despite using one-tenth as many parameters.

Keywords: Historical documents, Classical Chinese, OCR, HTR, vision–language models, long-tailed distribution, Dunhuang manuscripts

1. Introduction

The sinographic writing system has been used continuously for more than two millennia across a vast geographic area, giving rise to a written heritage so extensive that, as a Chinese saying goes, it could "make an ox sweat and fill a house up to the ceiling". Although the People's Republic of China was a pioneer in using automatic text recognition (ATR) to broaden access to this heritage[1], publicly available digital corpora remain limited in size, uneven in quality, and often redundant. Large-scale digitization therefore still depends on transcription systems that minimize post-correction cost while generalizing across collections.

Automatic transcription of historical sinographic documents is difficult not only because of the size of the character set and the strong visual similarity between classes (Shen et al., 2023), but also because of three structural properties. First, Chinese characters carry high information density: Shannon-entropy estimates suggest roughly $\sim$1.5 bits per Latin letter versus $\sim$10 bits per Chinese character, making a single character error potentially as consequential as an English word error. Under this view, a CER of 4% in Classical Chinese corresponds roughly to losing one word out of twenty-five in English. Second, character frequencies exhibit an extreme long-tail distribution, with roughly 20% of characters accounting for 95% of occurrences, which makes balanced coverage costly. Third, convention and standardization issues introduce additional brittleness: Unicode assigns distinct code points to glyph variants of the same underlying character (e.g., 黃 U+9EC3 vs. 黃 U+9EC4). Such duplications—numbering in the tens of thousands—complicate transcription conventions, hinder interoperability across corpora, and make evaluation unnecessarily fragile.

To address these challenges, we propose AnandaSky, a line-level transcription model for historical sinographic documents based on a vision–language architecture. The name combines Ananda—the disciple of the Buddha traditionally associated with the "encoding" of early Buddhist texts—and Sky, the opening character of the *Thousand Character Classic*, a text long used in premodern China to enumerate items. AnandaSky is designed to transcribe both historical prints and cursive handwriting.

Our contributions are as follows:

- We adapt the vision–language model paradigm to efficient line-level transcription of historical sinographic documents through a compact high-resolution architecture with global visual attention, 10 px patches, an uncompressed visual prefix, and variable-length attention.
- We introduce AnandaSky, a domain-specialized model trained on 4M line images, together with a systematic glyph normalization pipeline that reduces convention-induced brittleness across heterogeneous corpora.
- We show that this compact specialized model transfers robustly across collections, sets a new state of the art on MTHv2 (Ma et al., 2020), and outperforms much larger general-purpose baselines.

[1]This refers to the *Siku Quanshu* transcription project, launched in 1996, which digitized 3,461 works—about 800 million characters—in three years.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 311–321.

To support broader access to sinographic heritage, we release the AnandaSky model weights and inference code on Hugging Face.[2]

2. Related Work

Benchmarking efforts for historical Chinese document recognition include HDRC-CHINESE at ICDAR 2019 (Saini et al., 2019) and the CUHK Chinese OCR challenges in 2021 and 2022 (CUHK Library, 2021, 2022). A recurring pattern in these evaluations is the continued reliance on the classical OCR pipeline: character detection followed by isolated character classification. By contrast, modern handwritten text recognition (HTR) has largely moved toward line-level recognition with monotonic alignment methods such as CTC (Graves et al., 2006; Garrido-Muñoz et al., 2025), which avoid brittle segmentation and enable context-sensitive decoding.

More recently, ATR has increasingly been cast as a conditional generation problem in vision–language models (VLMs). TrOCR combines a pretrained Transformer-based vision encoder with a pretrained language decoder through cross-attention (Li et al., 2021), whereas DTrOCR explores a decoder-only alternative in which patch embeddings are supplied directly as a visual prefix, without a separate vision encoder (Fujitake, 2024). General-purpose VLMs such as Qwen2.5-VL (Bai et al., 2025) follow the now-standard design of a ViT-style visual encoder paired with an autoregressive LLM decoder conditioned on visual tokens, and perform strongly on modern text-in-image tasks. Historical sinographic documents, however, remain substantially out of distribution for such models: on historical Chinese benchmarks, even the best-performing general-purpose system, Gemini 2.5 Pro, achieves only 32.03% CER (Yu et al., 2025). This gap motivates domain-specialized VLMs that combine high-resolution visual encoding, convention-aware normalization, and large-scale historical training data.

3. Architecture

AnandaSky follows the Qwen2.5-VL paradigm, combining (i) a Vision Transformer (ViT) visual encoder and (ii) a pretrained autoregressive decoder. The full model contains 625.98M parameters. A line image is split into patches and encoded into a sequence of visual tokens, which are then projected into the decoder embedding space as a visual prefix that conditions autoregressive transcription generation.

Encoder component	Value
Patch size	10 px
Encoder embedding	768
Hidden size	768
Transformer layers	4
Attention heads	12
FFN size	2048
FFN activation	SwiGLU
Normalization	HybridNorm*

Table 1: Visual encoder configuration of AnandaSky.

Visual encoder. Table 1 summarizes the visual encoder configuration. The encoder is intentionally compact, with only four Transformer layers, hidden size 768, and 12 attention heads. We favor a shallow but high-resolution visual backbone because historical line transcription depends primarily on fine-grained glyph discrimination, while most long-range sequence modeling is handled by the pretrained decoder.

Unlike Qwen2.5-VL, which relies on local visual attention, we use global attention over the full line image. This choice is motivated by the sequential nature of text: visually relevant dependencies may extend across the entire line, for example when disambiguating confusable characters or compensating for missing strokes. We also reduce the patch size to 10 px (vs. 14 px in Qwen2.5-VL), which increases the visual sequence length but preserves crucial micro-details for cursive and degraded handwriting. Within the encoder, feed-forward blocks use SwiGLU MLPs, normalization follows the HybridNorm* strategy (Zhuo et al., 2025), and positional information is encoded with 2D Rotary Positional Encoding.

Visual prefix and decoder. While Qwen2.5-VL compresses visual tokens before injecting them into the decoder, we keep the full token sequence in order to preserve fine spatial information. The autoregressive decoder is initialized from Qwen3-0.6B[3]. It generates the transcription conditioned on the projected visual tokens, thereby integrating a pretrained language model directly into the decoding process. Following the multimodal positional encoding scheme of the Qwen-family models, the decoder uses MRoPE-Interleaved positional encoding (Huang et al., 2025).

Variable-length attention. Line images vary substantially in width and therefore in visual sequence length. To avoid excessive padding, we use variable-length attention in both the visual encoder and the autoregressive decoder. This reduces padding overhead and improves training efficiency.

[2] https://huggingface.co/badianeai/AnandaSky

[3] https://huggingface.co/Qwen/Qwen3-0.6B

4. Dataset

4.1. Main sinographic corpus

AnandaSky is trained on a large corpus comprising 66,607,288 character instances across 4,039,988 annotated lines. The line images come from historical sinographic documents produced between the 8th and the 20th century in China and Korea. The corpus includes 621,017 handwritten lines (about 15%) in cursive or semi-cursive scripts; among them, 337,696 lines come from Dunhuang manuscripts.

4.2. Transcription normalization

Because the transcriptions were produced by different teams using heterogeneous conventions, we apply a systematic normalization pipeline to reduce convention-induced variation. When differences correspond only to minor graphic variants—for example, 晋 U+664B vs. 晉 U+6649, along with associated forms such as 戬 U+622C vs. 戩 U+6229—we use a deterministic one-to-one mapping. For 2,267 characters for which no unique normalization can be defined, we instead perform context-sensitive disambiguation with a bidirectional encoder. A representative case is 后 U+540E, which may correspond either to the traditional form 後 U+5F8C or to a distinct lexical item meaning "queen." After normalization, the effective vocabulary is reduced to 16,770 distinct classes. A more detailed account of this normalization pipeline will be presented in a separate publication.

4.3. Multiscript pretraining corpus

For pretraining, we build a second dataset that adds other historical scripts (Latin manuscripts, French, Japanese, etc.) to the sinographic data. The full pretraining corpus totals roughly 10 million lines. It strengthens robustness to style and degradation and helps stabilize vision–language alignment at scale.

5. Training

Training proceeds in multiple phases to stabilize vision–language alignment while preserving the linguistic capabilities of the pretrained decoder. Throughout training, inputs are kept at near-native resolution: line images are fed without rescaling whenever possible, and only lines whose width exceeds 120 px are resized to this maximum width while preserving aspect ratio.

Phase 1: encoder pretraining. We first pretrain the visual encoder with a shallow autoregressive decoder. This stage adapts the visual backbone to the target document domain and provides a strong visual representation before learning the vision–language interface.

Phase 2: multimodal pretraining. We then train the multimodal model in two steps, following Qwen2.5-VL strategy: (i) we train only the adaptor for about 10% of total iterations to learn a stable projection into the decoder embedding space; (ii) we then jointly optimize the full model. To preserve linguistic representations in the decoder, we use differential learning rates: a higher learning rate for the visual encoder and adaptor, and a lower learning rate for the decoder. This limits catastrophic drift while allowing gradual adaptation to the visual signal.

Phase 3: fine-tuning. Finally, we fine-tune on the sinographic historical data to better capture paleographic specificities and the long-tail character distribution.

Dataset	# lines	CER (%)
Qing legal documents	142	4.89
Dunhuang manuscripts	841	1.38
MTHv2	8,051	0.92
Sibucongkan	1,146	0.43
Korean Anthologies	1,477	0.33

Table 2: Results on in-domain evaluation sets.

Dataset	# lines	CER (%)
ICDAR2019-HDRC	2,560	0.96
CUHK Challenge 2021	3,394	0.82
CUHK Challenge 2022	4,318	1.61

Table 3: Results on held-out benchmarks.

6. Evaluation

6.1. In-domain and external benchmarks

We evaluate AnandaSky on both in-domain test sets and external public benchmarks. The in-domain sets measure generalization to held-out documents from the same corpora as the training data, whereas the public benchmarks provide a stricter cross-dataset evaluation, since no samples from these datasets were used during training.

To account for heterogeneous transcription conventions, we normalize all ground-truth transcriptions using the same normalization pipeline as in training. Decoding is performed with beam search (beam size 4), and performance is reported in terms of character error rate (CER), defined as the Levenshtein distance normalized by the reference length and expressed as a percentage.

Reference	CER (%)
Ma et al. 2020	4.48
Shi et al. 2017	3.06
Huang et al. 2021	2.58
Huang et al. 2023	2.11
AnandaSky (ours)	**0.92**

Table 4: Comparison with previously reported results on MTHv2. Best result in bold.

As shown in Tables 2–3, AnandaSky delivers consistently strong performance on both in-domain and external evaluations. CER falls below 1% on five of the eight datasets, including three in-domain sets and two held-out benchmarks. On MTHv2 (Ma et al., 2020), AnandaSky achieves 0.92% CER (Table 4), improving on the best previously reported result of 2.11% by 1.19 absolute points. Performance on ICDAR2019-HDRC (Saini et al., 2019) and CUHK Challenge 2021 (CUHK Library, 2021) is particularly noteworthy, since these datasets consist of family records and Taoist texts, respectively, document types that are absent from the training distribution. Taken together, these results indicate that the model generalizes robustly across linguistic and documentary domains.

Further error analysis is provided in Appendix A.2. Figure 2 shows that errors are highly concentrated near zero across all in-domain datasets, confirming that most lines are transcribed perfectly or with only minor deviations. The aggregate CER is therefore driven by a relatively small upper tail of difficult cases. To examine this tail while limiting selection bias, Figures 3–7 present stratified random samples from each dataset: panels A and B are drawn from lines above the 90th percentile of line-level CER, while panels C and D are drawn from the 80th–90th percentile band. Manual inspection suggests that many of the highest-CER cases are associated with image degradation (e.g., Figure 3, panels A–B), normalization mismatches (e.g., Figure 4), or apparent inconsistencies in the reference transcriptions (e.g., Figures 6–7), rather than systematic recognition failures.

More generally, these observations highlight that quantitative evaluation is only as reliable as the reference transcriptions on which it is based. In particular, while the reported CER suggests that manuscript transcription remains more difficult than print transcription, the qualitative analysis of the Dunhuang test set (Figure 4) points to a more nuanced picture: a substantial fraction of the apparent errors seem to reflect annotation noise, normalization mismatches, or ambiguous ground truth rather than genuine reading failures. The reported CER for this dataset may therefore overestimate the true transcription error, and actual performance may be closer to the sub-1% regime than the raw score alone suggests.

6.2. EvaHan 2026 shared-task results

EvaHan 2026 is a shared task on multimodal methods for ancient Chinese document understanding, comprising line-level transcription for printed materials (Task A), line-level transcription for manuscripts (Task C), and document segmentation (Task B). In this section, we report our official open-track results for the transcription tasks. Task B is discussed separately in Appendix A.1.

For Tasks A and C, we submitted an open-track system obtained by fully fine-tuning AnandaSky on the official training data. We chose full fine-tuning because EvaHan adopts transcription conventions that differ from our own; without adaptation, AnandaSky would be penalized by systematic convention mismatches rather than genuine recognition errors.

Table 5 reports the official results communicated by the organizing committee for Tasks A and C. Despite using only 626M parameters—more than 10$\times$ fewer than the 7B-parameter Qwen2.5-VL baseline—AnandaSky-ft consistently outperforms the official baseline across all reported metrics, reducing CER by 5.2% on Task A and by 12.1% on Task C. The gains are particularly notable because they are obtained after adapting a much smaller domain-specialized model to the organizers' conventions.

We note, however, that the official EvaHan CER is substantially higher than the CER obtained by AnandaSky on the other benchmarks considered in this paper. Because the EvaHan test set is not publicly available, we can only propose plausible explanations. First, the official evaluation appears to rely on macro averaging, whereas all other scores reported in this paper use micro averaging. Second, the EvaHan test distribution may differ from the development data, and full fine-tuning on the organizers' training set may trade some cross-collection robustness for improved alignment with task-specific annotation conventions.

7. Ablation

We ablate the encoder patch size to quantify the impact of visual granularity on transcription accuracy. For each patch size, we rerun the full training pipeline. To ensure a fair comparison, we adjust the batch size so that each configuration is trained on the same total number of images, while keeping the remaining hyperparameters unchanged.

Overall performance improves only moderately as the patch size decreases (14 px: CER=1.96%;

Task	System	CER	NED	F1	Overall
A	Qwen2.5-VL 7B (baseline)	0.0618	0.0613	0.9430	0.9397
A	**AnandaSky-ft**	**0.0586**	**0.0564**	**0.9479**	**0.9438**
$A_{\text{no-var}}$	Qwen2.5-VL 7B (baseline)	0.0685	0.0679	0.9364	0.9331
$A_{\text{no-var}}$	**AnandaSky-ft**	**0.0639**	**0.0617**	**0.9426**	**0.9385**
C	Qwen2.5-VL 7B (baseline)	0.0920	0.0919	0.9099	0.9086
C	**AnandaSky-ft**	**0.0809**	**0.0809**	**0.9200**	**0.9194**
$C_{\text{no-var}}$	Qwen2.5-VL 7B (baseline)	0.1066	0.1065	0.8953	0.8940
$C_{\text{no-var}}$	**AnandaSky-ft**	**0.1007**	**0.1007**	**0.9002**	**0.8996**

Table 5: Official EvaHan 2026 open-track results for Tasks A and C, comparing the official Qwen2.5-VL 7B baseline (instruction-tuned) with AnandaSky-ft, i.e., AnandaSky fully fine-tuned on the EvaHan training data. Rows tagged $_{\text{no-var}}$ correspond to the stricter *excluding-variants* evaluation; untagged rows use the *including-variants* setting. Best results per task/setting are shown in **bold**.

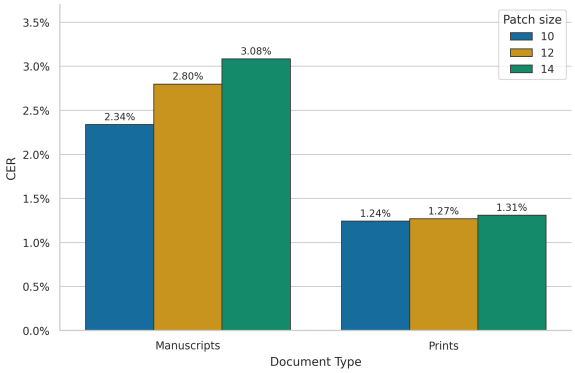

Figure 1: CER by document type as a function of the encoder patch size.

12 px: CER=1.82%; 10 px: CER=1.63%). However, stratifying results by document type reveals a clear and consistent trend. As displayed in Figure 1, finer patches yield substantially larger gains on manuscripts than on prints. On manuscripts, reducing the patch size from 14 px to 10 px lowers CER from 3.08% to 2.34%, highlighting the importance of capturing fine-grained stroke details for cursive and degraded handwriting. In contrast, the effect on printed documents is smaller, with CER improving from 1.31% (14 px) to 1.24% (10 px). These results support our choice of a 10 px patch size: the increased visual sequence length primarily benefits the most challenging regime while leaving printed-text performance largely unchanged.

8. Conclusion

We introduced AnandaSky, a vision–language model for line-level transcription of historical sinographic documents. By adapting the VLM paradigm to this setting—through a compact high-resolution visual encoder with global attention, 10 px patches, an uncompressed visual prefix, and efficient variable-length attention—and by training at scale with convention-aware transcription normalization, we obtain strong results on both in-domain and held-out benchmarks. AnandaSky sets a new state of the art on MTHv2, reaches sub-1% CER on five of eight evaluation datasets, and transfers robustly across unseen collections.

Beyond benchmark performance, our results suggest that the field is approaching a regime in which quantitative evaluation alone is no longer sufficient. As error rates become very low, aggregate CER increasingly reflects not only model mistakes but also annotation noise, normalization mismatches, and inconsistencies in reference transcriptions. In particular, while the reported scores remain somewhat higher on manuscripts than on prints, the qualitative analysis of the Dunhuang test set suggests that this gap is smaller than the raw CER indicates, and that actual manuscript transcription performance may be closer to the low-error regime reached on printed materials.

From a practical perspective, this means that large-scale transcription of historical sinographic corpora can increasingly shift from labor-intensive manual correction toward lightweight verification. More broadly, our findings show that compact domain-specialized VLMs can be highly effective for historical document transcription, and provide a strong foundation for future systems that integrate transcription with layout analysis, retrieval, and richer document understanding.

9. Limitations

Our work has several limitations. First, the competition setting constrained the amount of exploration we could perform in terms of scale. We did not systematically study the effect of longer training schedules, larger visual backbones, or larger autoregressive decoders within the EvaHan timeline. Since both VLM performance and robustness often improve with compute and model capacity, a

more exhaustive scaling study (parameters, data, and training duration) is left for future work.

Second, while AnandaSky generalizes well across the printed and manuscript benchmarks considered in this paper, its robustness is not uniform across all sinographic document domains. In our experiments, we observed strong performance on Qing dynasty legal documents, but noticeably weaker results on similar legal materials produced in Indonesia (Dean et al., 2025). This suggests that domain shifts in writing conventions or language can still degrade performance. Addressing such cross-region and cross-collection shifts will likely require additional targeted data, improved normalization of conventions, or explicit domain adaptation strategies.

Finally, our model operates at the line level and therefore assumes that reliable line extraction is available. Deploying AnandaSky in end-to-end digitization pipelines requires upstream layout analysis and line detection; errors in these components can propagate to transcription quality. Integrating robust layout modeling and reading-order prediction with line-level transcription remains an important direction for future work.

10. Acknowledgements

We thank the organizers of the 2021 and 2022 Chinese OCR Challenges hosted by the CUHK Library for granting access to the competition data. This work was supported by the French National Research Agency (ANR) under the Investissements d'avenir programme (ANR-21-ESRE-0005, EquipEx Biblissima+), and by the French government through the France 2030 investment plan managed by ANR, as part of the Initiative of Excellence Université Côte d'Azur (ANR-15-IDEX-01). This work also benefited from access to the HPC resources of IDRIS under GENCI allocations 2026-AD011014973R2 and 2026-AD011017524R1. This research was also funded in part by ANR under project ANR-24-CE27-4500-03.

11. Bibliographical References

Shuai Bai et al. 2025. Qwen2.5-VL technical report. arXiv preprint.

CUHK Library. 2021. 2021 chinese classic text OCR challenge. Webpage. Accessed 2026-02-28.

CUHK Library. 2022. Chinese classic text OCR challenge 2022. Webpage. Accessed 2026-02-28.

Kenneth Dean, Maria De Iorio, Alexander Mozdzen, Zhang Qiao, Yu Kang, and Andrew Harris. 2025. Unveiling country trade: Insights from 19th century batavia Kong Koan records via OCR and NLP analysis. Webpage. Accessed 2026-02-28.

Masato Fujitake. 2024. DTrOCR: Decoder-only transformer for optical character recognition. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8010–8020.

Carlos Garrido-Muñoz, Antonio Rios-Vila, and Jorge Calvo-Zaragoza. 2025. Handwritten text recognition: A survey. arXiv preprint.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Jiarong Huang, Dezhi Peng, Hongliang Li, Hao Ni, and Lianwen Jin. 2023. SegCTC: Offline handwritten chinese text recognition via better fusion between explicit and implicit segmentation. In *Document Analysis and Recognition – ICDAR 2023*, pages 332–349, Cham. Springer Nature Switzerland.

Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. 2025. Revisiting multimodal positional encoding in vision-language models.

Yuhao Huang, Lianwen Jin, and Dezhi Peng. 2021. Zero-shot chinese text recognition via matching class embedding. In *Document Analysis and Recognition – ICDAR 2021*, pages 127–141, Cham. Springer International Publishing.

Minghao Li et al. 2021. TrOCR: Transformer-based optical character recognition with pre-trained models. arXiv preprint.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. 2020. Joint layout analysis, character detection and recognition for historical document digitization. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 31–36, Los Alamitos, CA, USA. IEEE Computer Society.

Rajkumar Saini, Derek Dobson, Jon Morrey, Marcus Liwicki, and Foteini Simistira Liwicki. 2019. ICDAR 2019 historical document reading challenge on large structured chinese family records. In *Proceedings of the International Conference*

on Document Analysis and Recognition (ICDAR), pages 1499–1504.

Lu Shen, Bidong Chen, Jianjing Wei, Hui Xu, Su-Kit Tang, and Silvia Mirri. 2023. The challenges of recognizing offline handwritten chinese: A technical review. *Applied Sciences*, 13(6):3500.

Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Haiyang Yu, Yuchuan Wu, Fan Shi, Lei Liao, Jinghui Lu, Xiaodong Ge, Han Wang, Minghan Zhuo, Xuecheng Wu, Xiang Fei, Hao Feng, Guozhi Tang, An-Lan Wang, Hanshen Zhu, Yangfan He, Quanhuan Liang, Liyuan Meng, Chao Feng, Can Huang, Jingqun Tang, and Bin Li. 2025. Benchmarking vision-language models on chinese ancient documents: From OCR to knowledge reasoning.

Zhijian Zhuo, Yutao Zeng, Ya Wang, Sijun Zhang, Jian Yang, Xiaoqing Li, Xun Zhou, and Jinwen Ma. 2025. Hybridnorm: Towards stable and efficient transformer training via hybrid normalization.

A. Appendices

A.1. EvaHan 2026 Task B Results

For this track, we used a YOLO26 model with an input resolution of 960 px. The official Task B results are reported in Table 6. Our system substantially outperforms the official Qwen2.5-VL 7B baseline, improving mAP@[.5:.95] from 0.2006 to 0.4389, Micro F1 from 0.0513 to 0.7232, Macro F1 from 0.1530 to 0.6657, and Avg Match IoU from 0.6600 to 0.8114. We nevertheless interpret the absolute Task B scores with caution. Manual inspection of the available annotations suggests that this task is affected by non-negligible annotation noise and matching inconsistencies, likely to a greater extent than Tasks A and C. We therefore regard the relative improvement over the official baseline as more informative than the absolute score level.

Task	System	mAP@[.5:.95]	Micro F1	Macro F1	Avg Match IoU
B	Qwen2.5-VL 7B (baseline)	0.2006	0.0513	0.1530	0.6600
B	**Ours**	**0.4389**	**0.7232**	**0.6657**	**0.8114**

Table 6: Official EvaHan 2026 open-track results for Task B, comparing the official Qwen2.5-VL 7B baseline (instruction-tuned) with our YOLO26-based system. Best results are shown in **bold**.

A.2. Qualitative error analysis on in-domain datasets

Figure 2 shows the empirical cumulative distribution of line-level CER on the in-domain evaluation sets. In all datasets, most lines cluster near zero CER, indicating that errors are concentrated in a relatively small upper tail of difficult cases. To examine this tail, Figures 3–7 present stratified random samples from each dataset: panels A and B are drawn from lines above the 90th percentile of line-level CER, while panels C and D are drawn from the 80th–90th percentile band. In each panel, the text shown on the right is the transcription generated by the model during evaluation, and red characters indicate mismatches between the generated transcription and the ground truth.

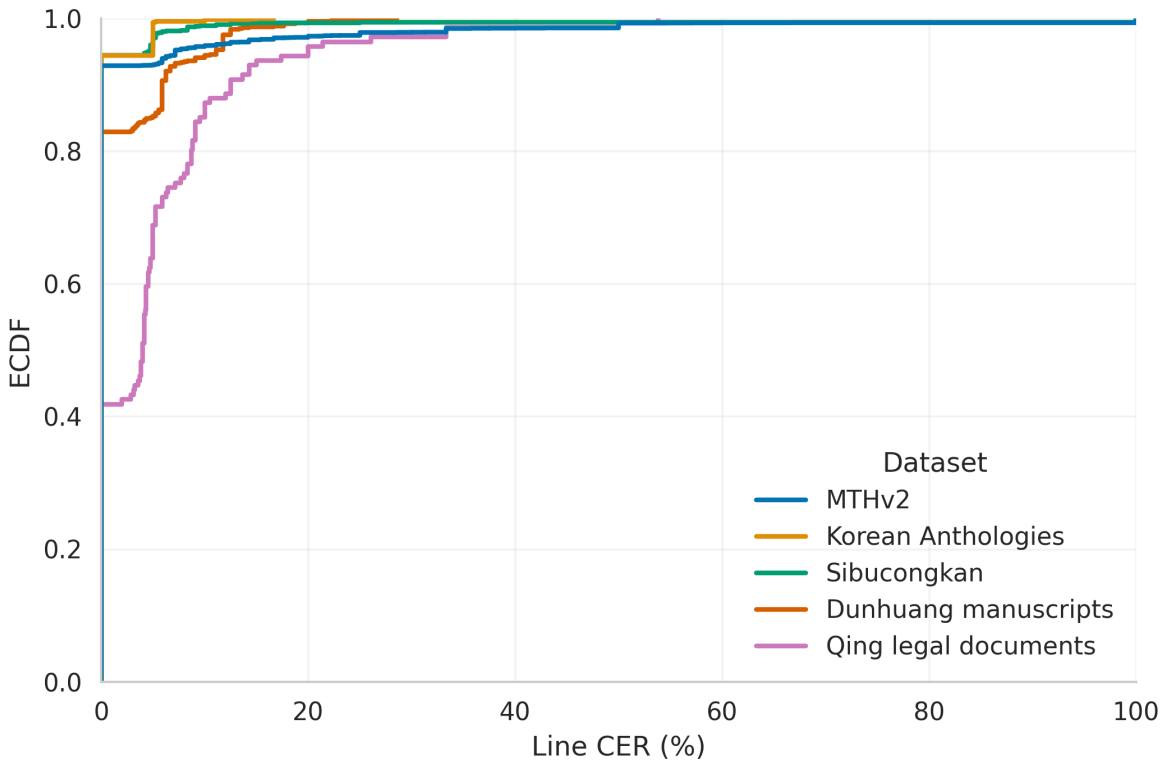

Figure 2: ECDF of line-level CER on the in-domain evaluation sets. Most lines have zero or very low CER, with aggregate error driven by a relatively small upper tail of difficult examples.

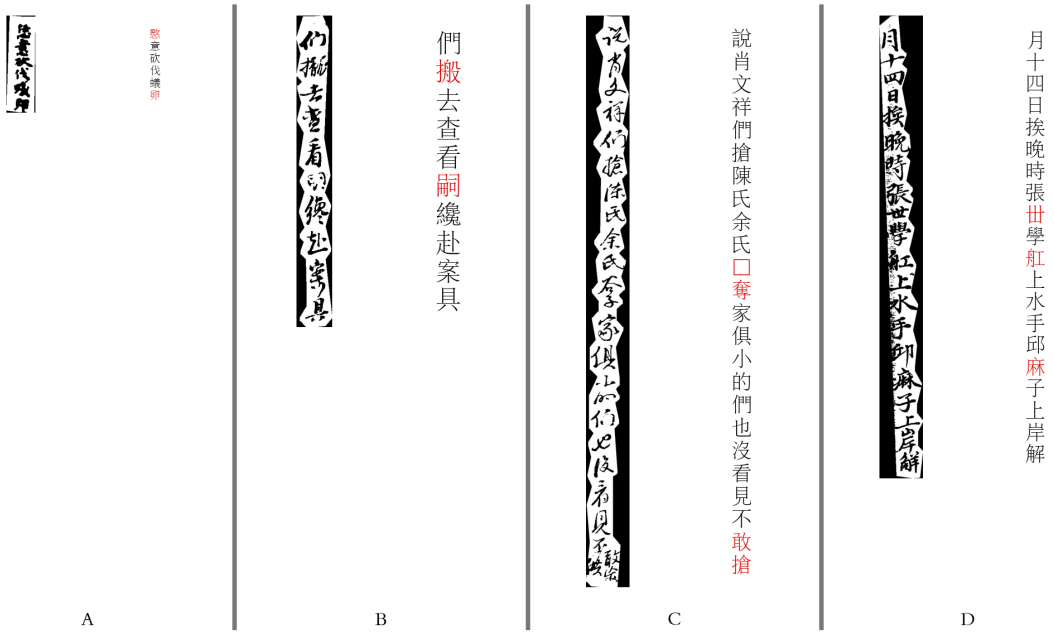

Figure 3: Upper-tail error examples from the Qing legal documents dataset.

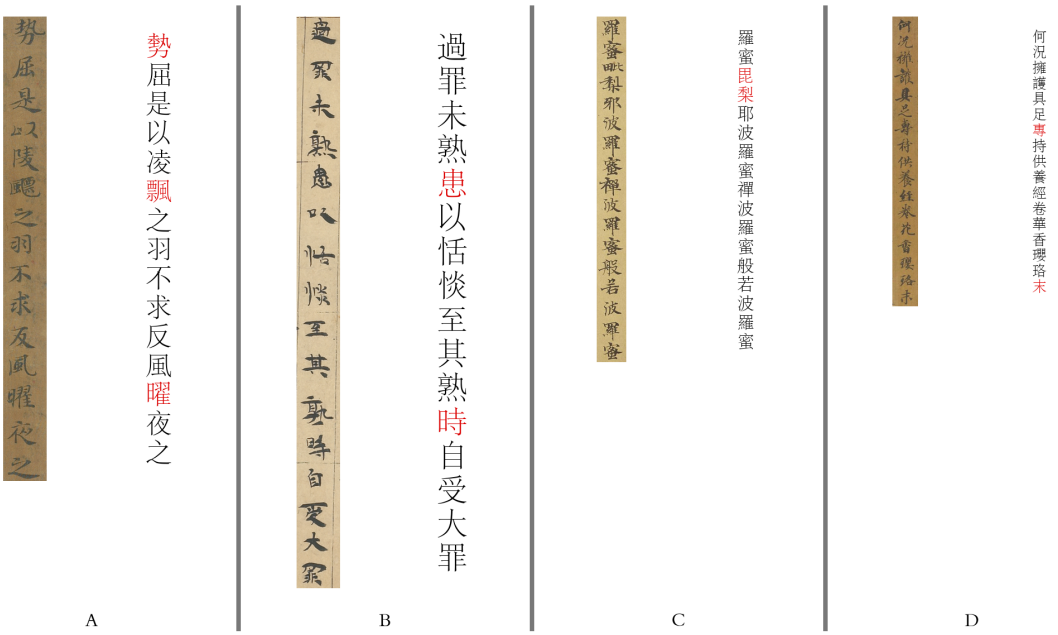

Figure 4: Upper-tail error examples from the Dunhuang manuscripts dataset.

Figure 5: Upper-tail error examples from the MTHv2 dataset.

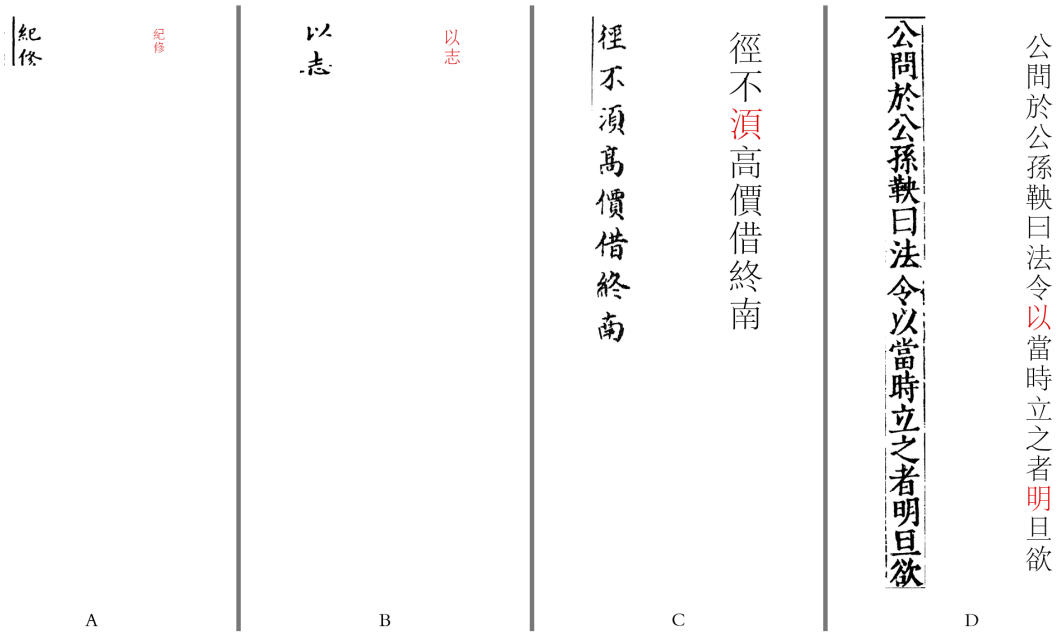

Figure 6: Upper-tail error examples from the *Sibu congkan* dataset.

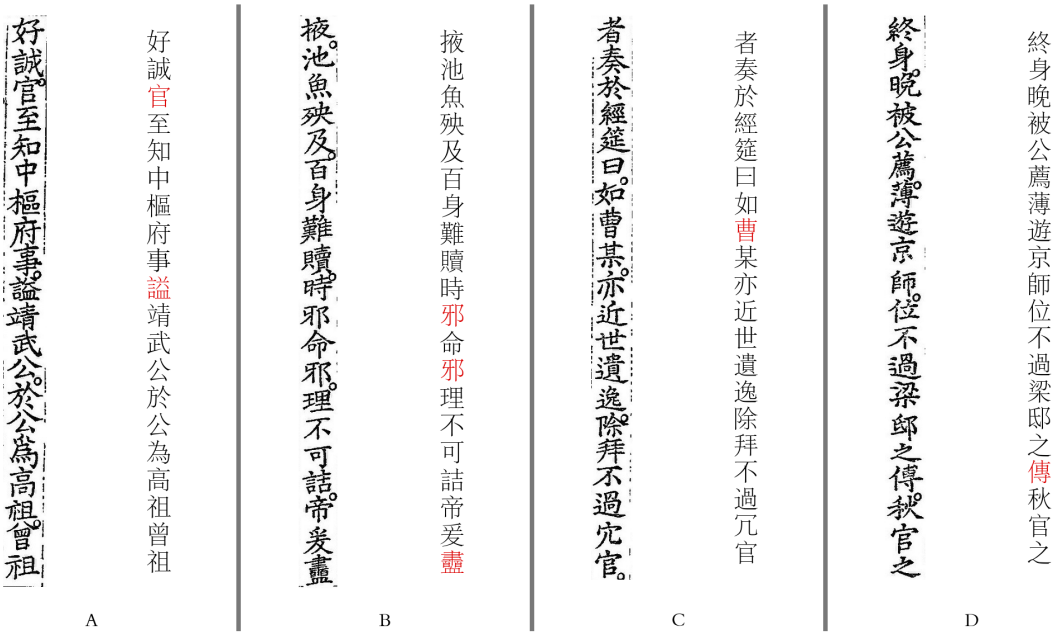

Figure 7: Upper-tail error examples from the Korean Anthologies dataset.