Recontextualizing Revitalization: A Mixed Media Approach to Reviving the Nüshu Language

Ivory Yang **Xiaobo Guo** **Yuxin Wang** **Hefan Zhang**
Yaning Jia **William Dinauer** **Soroush Vosoughi**
Department of Computer Science, Dartmouth College
{Ivory.Yang.GR, Soroush.Vosoughi}@dartmouth.edu

Abstract

Nüshu is an endangered language from Jiangyong County, China, and the world's only known writing system created and used exclusively by women. Recent Natural Language Processing (NLP) work has digitized small Nüshu-Chinese corpora, but the script remains computationally inaccessible due to its handwritten, mixed-media form and dearth of multimodal resources. We address this gap with two novel datasets: *NüshuVision*, an image corpus of 500 rendered sentences in traditional vertical, right-to-left orthography, and *NüshuStrokes*, the first sequential handwriting recordings of all 397 Unicode Nüshu characters by an expert calligrapher. Evaluating five state-of-the-art Chinese Optical Character Recognition (OCR) systems on *NüshuVision* shows that all fail entirely, each yielding a Character Error Rate (CER) of 1.0. Fine-tuning Microsoft's TrOCR on NüshuVision lowers CER to 0.67, a modest yet meaningful improvement. These contributions establish the first multimodal foundation for Nüshu revitalization and offer a culturally grounded framework for language preservation.

1 Introduction

Nüshu is an endangered language from Jiangyong County, Hunan, China, and the world's only known writing system developed and used exclusively by women (Zhao, 1998). It emerged as a private means of expression in a patriarchal society where women were largely excluded from formal education (Li, 2024). With the passing of the last native speaker in 2004 (Zuo and Sirisuk, 2024), UNESCO classified Nüshu as critically endangered (Liu, 2018); today, literacy survives only among a small group of scholars and revivalists (Hu, 2022). Recent NLP work has introduced digitized Nüshu-Chinese corpora (Yang et al., 2025b), but these efforts focus narrowly on token-level translation. In practice, **most surviving Nüshu materials are handwritten on mixed media**, and lack standardized transcriptions

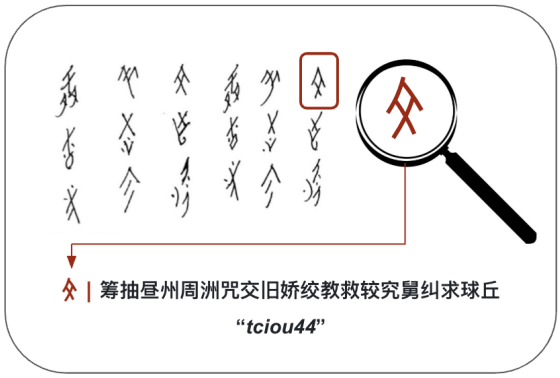


Figure 1: OCR pipeline identifying Nüshu[A.3] characters from handwritten images, mapping each to its Unicode, Chinese equivalent, and Jiangyong pronunciation.

(Lim and Bahauddin, 2025). This creates a fundamental bottleneck: before deeper linguistic analysis or translation can proceed, Nüshu must first be made machine-readable across visual (e.g., OCR) and temporal (e.g., stroke sequence) modalities.

To address this, we present a multimodal approach targeting the mixed media nature of Nüshu, centered on two new datasets. (1) *NüshuVision*: OCR-compatible corpus of 500 rendered sentence images in traditional vertical, right-to-left orthography. (2) *NüshuStrokes*: Sequential handwriting recordings of all 397 officially encoded Unicode Nüshu characters, enabling fine-grained modeling of stroke dynamics. Benchmarking five state-of-the-art Chinese OCR systems on *NüshuVision* yields CER of 100%, highlighting inadequacy of current tools, but fine-tuning Microsoft's TrOCR model (Li et al., 2023) improves recognition accuracy by over 30%. These contributions offer a reproducible blueprint for interdisciplinary revitalization of endangered languages such as Nüshu[1].

[1]This work embodies EMNLP 2025's special theme of *Interdisciplinary Recontextualization of NLP* by advancing multimodal approaches for language preservation - especially poignant given its venue of China, the birthplace of Nüshu.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12431–12440
November 4-9, 2025

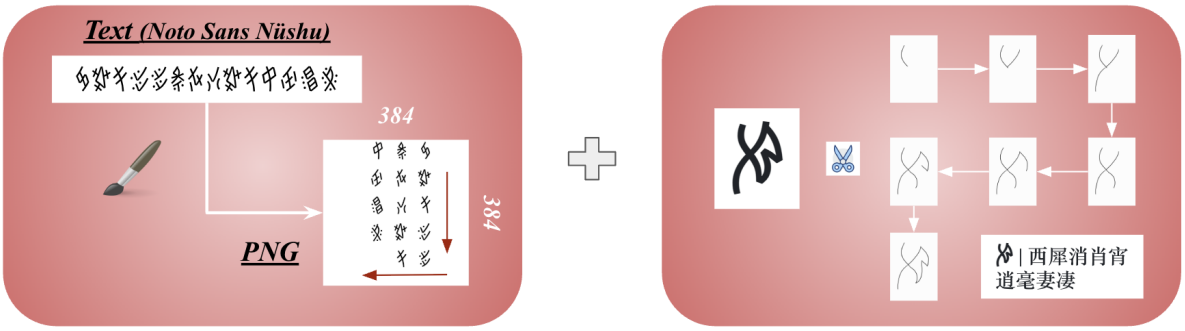


Figure 2: Visualization of the *NüshuVision* and *NüshuStrokes* dataset components and creation process.

2 Related Work

In recent years, Nüshu has gained cultural visibility, notably in the Hollywood film *Snow Flower and the Secret Fan* (Liu, 2024), yet linguistic research remains sparse. The most comprehensive resource, *A Compendium of Chinese Nüshu* (Zhao, 1992), offers a modest collection of scanned calligraphy with Chinese translations, while *Heroines of Jiangyong* (Chiu, 2012) includes English translations without formal alignment. Recent NLP work has cited Nüshu as a creative case for language emergence (Tang et al., 2023; Sun et al., 2023), and Yang et al. (2025b) introduced the first digitized Nüshu–Chinese dataset with a synthetic expansion framework. However, synthetic data for endangered languages remains controversial (Low et al., 2022), and multimodal modeling is largely absent, despite the fact that historical Nüshu survives primarily as handwritten calligraphy on mixed media (Zuo and Sirisuk, 2024). Addressing this modality gap is a critical step toward computational revitalization. A more detailed discussion of Nüshu's historical and linguistic background is included in Appendix A.

3 Datasets

To advance computational modeling of Nüshu, we introduce two complementary datasets: (1) *NüshuVision* designed for OCR, and (2) *NüshuStrokes* for sequential stroke-order on a character level, depicted in Figure 2.

3.1 *NüshuVision*: OCR-Compatible Images

Yang et al. (2025b) introduced the first digitized Nüshu–Chinese parallel corpus of 500 expert-validated sentences. While their work concentrated on linguistic analysis using text-based NLP techniques, our goal centers on OCR for Nüshu. To facilitate our objective, we constructed a corresponding image dataset, ***NüshuVision***, by rendering each sentence in a style faithful to traditional Nüshu orthography (Galambos, 2017), in which characters are output in vertical right-to-left script orientation (Schmandt-Besserat and Erard, 2009). Each character was drawn using the *NotoSansNushu-Regular.ttf* Unicode font (Google, 2020), with a base font size of 30 and 25-pixel vertical spacing. For longer sentences, font size was adaptively reduced to preserve legibility and spatial fit. Characters were grouped into columns based on vertical height constraints, reversed to reflect traditional column ordering, and horizontally padded for balance. Each rendered layout was then resized and centered within a fixed 384×384 pixel canvas using Lanczos resampling (Madhukar and Narendra, 2013) to ensure compatibility with TrOCR's (Li et al., 2023) input specifications.

3.2 *NüshuStrokes*: Sequential Handwriting

In logographic[2] writing systems such as Chinese (Ho and Bryant, 1997), the order in which a character's strokes are written is not arbitrary (Zhang, 2014). Stroke order carries semantic, stylistic, and structural significance (Chen, 1996), affecting both human legibility and machine interpretation (Fan and Lin, 1999). Although Nüshu is a syllabic[3] script (Congrong, 2024), stroke order remains deeply significant (Wang and Zhu, 2010). This is especially critical for handwriting recognition systems, where subtle temporal patterns can guide more accurate decoding (Sharma and

[2]Each character represents whole words or morphemes (meaningful units of language) rather than individual syllables.

[3]Each character corresponds to a spoken syllable rather than a morpheme or word.

Jayagopi, 2021). In response, we collaborated with a professional ancient Chinese calligrapher with over 11 years of formal training, who **handwrote all 397 officially encoded Unicode Nüshu characters** (The Unicode Consortium, 2017).

To capture the temporal dynamics of stroke formation in Nüshu writing, we recorded each character's writing process as an MP4 video using the Ophaya 3-in-1 Smartpen Set (oph, 2025). The average duration of the 397 MP4 videos was 3.94 seconds. We then extracted key frames from each video to isolate the sequential structure of individual strokes. To identify significant transitions, we computed the mean squared error (MSE) between consecutive grayscale frames. When the MSE between adjacent frames remained below a threshold of 0.1 for more than one frame, we classified the sequence as static and extracted the earliest frame from that segment. This process ensured we preserved the visual onset of each discrete stroke. The MSE was computed as follows:

$$\text{MSE} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(I_A(i,j) - I_B(i,j)\right)^2 \quad (1)$$

where $I_A(i,j)$ and $I_B(i,j)$ represent the pixel intensities of two consecutive grayscale frames at position (i,j), and $m \times n$ is the resolution of each frame. The motivation, potential and planned use for this dataset is expanded upon in Appendix C. ***NüshuStrokes* is the first-of-its-kind dataset capturing temporal dynamics of handwritten Nüshu, offering a foundational resource for interdisciplinary research and revitalization**

4 Benchmark on Chinese OCR Models

1. Performance Given Nüshu's historical and visual affinities with Chinese script, particularly in its use of vertical layout and stroke-based composition (Wang, 2020), it is reasonable to consider whether modern Chinese OCR systems might generalize to this endangered script. To explore this, we evaluated five state-of-the-art Chinese OCR engines on our *NüshuVision* dataset: PaddleOCR (v3.6) (Li et al., 2022), Tesseract (v5.4, with `-oem 3` and both `chisim/chitra` models) (Smith, 2007), EasyOCR (v1.7.2) (JaidedAI, 2020), Google Cloud Vision API (latest) (Google Cloud, 2023), and Tencent Cloud OCR (Tencent Cloud, 2023). Each engine was tested in both simplified and traditional Chinese modes. As shown in Table 1, all systems failed to recognize Nüshu characters, yielding a Character Error Rate (CER) of 1.00, indicating that 0% of characters were predicted correctly. CER ranges from 0 to 1 (0 denotes perfect recognition, 1 indicates complete failure) and was computed as follows:

OCR Engine	1 - CER (Simp.)	1 - CER (Trad.)
PaddleOCR	0.0	0.0
Tesseract 5.4	0.0	0.0
EasyOCR 1.7.2	0.0	0.0
Google Cloud Vision	0.0	0.0
Tencent Cloud OCR	0.0	0.0
TrOCR Finetuned	**0.33**	**0.33**

Table 1: Nüshu detection accuracy (1 - Character Error Rate) for five Chinese OCR engines and our TrOCR model on *NüshuVision* dataset, evaluated in both simplified (Simp.) and traditional (Trad.) Chinese modes.

$$\text{CER} = \frac{S + D + I}{N} \quad (2)$$

where S = number of substitutions, D = number of deletions, I = number of insertions, N = total number of characters in the reference (ground truth) sequence. These results highlight that despite superficial visual similarities, Nüshu's underlying structure and semantics are distinct enough to render Chinese OCR models entirely ineffective, necessitating script-specific solutions.

2. Cross-Script Recognition This result, while disappointing, is ultimately unsurprising: Chinese OCR systems are trained on distinct orthographic conventions and semantic mappings that diverge significantly from those of Nüshu. Nonetheless, occasional glimmers of cross-script recognition emerge in specific cases where individual Nüshu characters bear strong visual resemblance to common Chinese logograms, particularly numerals like "one", "two", or pictographic forms such as "person". In such instances, the OCR engine often detects the Chinese mapping of the Nüshu character, rather than the Nüshu character itself. While these errors highlight a fundamental limitation in the models' ability to distinguish Nüshu, they also hint at latent potential: *shared visual morphology between Nüshu and Chinese characters may offer indirect signals that could be systematically leveraged for weak supervision or pretraining, serving as a potential bridge in developing future Nüshu-specific OCR models*.

5 Finetuning TrOCR on Nüshu

1. Set-Up To create a script-specific OCR model for Nüshu, we fine-tuned Microsoft's TrOCR `trocr-base-stage1` encoder–decoder model using our *NüshuVision* dataset, which was randomly split into 400 training and 100 test samples. Each image was resized to 384×384 pixels and normalized via the built-in TrOCRProcessor, while target Nüshu sequences were tokenized with a maximum length of 128. We trained the model for 2000 epochs using the Hugging Face `Seq2SeqTrainer` on a single GPU, with learning rate of 3×10^{-5}, batch size = 16, and weight decay of 0.01 using the AdamW optimizer. Generation during evaluation employed beam search of 4 beams, length penalty of 2.0, and no-repeat n-gram size of 3.

2. Evaluation We evaluated the final saved checkpoint on a held-out test set of 100 images. The model achieved CER of 0.6735 (scale of 0 to 1), meaning roughly 33% of Nüshu characters were predicted correctly. While the model's performance is still far from human-level accuracy, it represents a significant step forward compared to five state-of-the-art Chinese OCR systems, all of which scored a CER of 1.00 - failing entirely to recognize any Nüshu characters. *These results show that even with limited authentic training data, general-purpose vision–language models like TrOCR can begin to learn and adapt to previously unsupported scripts such as Nüshu.*

6 Discussion

6.1 Performance vs Integrity of Data

While our fine-tuned TrOCR model achieves a modest CER of 0.67, these results must be contextualized within the unique linguistic and ethical landscape of Nüshu. As the only known language developed and used exclusively by women, and a recognized form of intangible cultural heritage, Nüshu carries immense historical and symbolic significance. **Its revitalization cannot, and should not, be reduced to outperforming standard benchmarks**. The challenge here is not just technical but epistemological: we are modeling a script with no living fluent users, extremely limited data, and few means of validation. Broader discussion on this topic, along with some reflection on revitalization work for endangered languages (Yang et al., 2025a), is featured in Appendix D.

In theory, synthetic data augmentation can inflate performance metrics. However, researchers have cautioned against this in particular low-resource language contexts (Anastasopoulos et al., 2020; Bird, 2024; Yang et al., 2025d). In Nüshu's case, where expert validation is scarce (Bird, 2020), higher benchmark scores through synthetic data risks codifying stylistic noise or imagined usage as linguistic fact (Chen et al., 2023; Alvarez et al., 2025; Yang et al., 2025c). We take a more cautious, culturally responsible approach; by showing that TrOCR can begin to learn Nüshu from just 400 authentic examples, we establish a meaningful lower bound for future work. The model's moderate success is not a limitation but a signal: Nüshu is learnable, but only through careful grounding in real data. As this year's EMNLP theme emphasizes, **benchmarks matter most when they reflect real-world complexity**. The difficulty here is not noise to eliminate, but a feature of the task itself.

6.2 Future Work

We are currently augmenting existing digitized Nüshu corpora with newly transcribed, expert-validated materials, and collecting oral recordings of the Jiangyong dialect to align with corresponding Nüshu characters. We also plan to apply stroke-level masking with the *NüshuStrokes* dataset (Appendix C), enabling models to learn which temporal features contribute most to recognition. This work continues in close collaboration with Nüshu-literate scholars and ancient Chinese calligraphers (Appendix B), whose domain expertise helps ensure both linguistic and cultural fidelity.

7 Conclusion

We present the first multimodal computational framework for Nüshu, combining rendered sentence images with sequential stroke recordings to support OCR and script modeling. Our evaluations show that existing Chinese OCR systems fail entirely on Nüshu, while fine-tuning TrOCR on just 400 authentic examples yields a measurable improvement. Rather than pursuing inflated scores through synthetic data, we prioritize cultural integrity and verifiable learning with expert-validated materials. Grounded in real data and cautious modeling, this work offers a reproducible foundation for endangered script revitalization, bridging traditional calligraphy and modern NLP, and exemplifying how interdisciplinary approaches can meaningfully extend the reach of language technologies.

Limitations

While our work lays the first multimodal foundation for Nüshu OCR, it is naturally shaped by the realities of modeling an endangered language with no surviving native users. Nonetheless, our research integrates and benefits from contributors with deep engagement in Nüshu scholarship and traditional calligraphy, ensuring that both curation and evaluation reflect informed domain expertise. Given that our model was trained on 400 authentic examples and tested on 100, it faces challenges typical of data-scarce settings. Moreover, our *NüshuStrokes* dataset was created by a single calligrapher. Nüshu script style varies by calligrapher, as does the writing direction; thus this dataset does not represent the full range of Nüshu script. In low-resource endangered script OCR, it is standard to begin with the most consistent subset to establish baseline. At the very start, under-representation is better than non-representation (current status) or over-representation. Layout generalization is absolutely essential for future integration, but depends on having a functioning core recognizer first, which is our paper's contribution.

Ethics

In the spirit of transparent and responsible research, we have made our code, the complete *NüshuVision* dataset, and the complete *NüshuStrokes* dataset publicly available at (https://github.com/ivoryayang/NushuMultimodal). Given Nüshu's cultural sensitivity and endangered status, all data (rendered images, handwriting recordings) was produced in close collaboration with Nüshu-literate scholars and trained calligraphers to ensure authenticity. We intentionally avoid synthetic data generation, which risks distorting the linguistic and stylistic reality of the script. All released materials are provided under open-access terms for educational and preservation-focused use. Our goal is not commercialization but cultural stewardship: to support Nüshu's survival through transparent, collaborative research that honors its historical and symbolic significance.

Acknowledgment

This work was partially supported by the CompX Faculty Grant from the *Neukom Institute for Computational Science* at Dartmouth College.

References

2025. Ophaya 3-in-1 Smartpen Set (Model APNX-70103). https://www.amazon.com/dp/B0DP3P3SRD. Accessed: 2025-05-18.

Jesus Alvarez, Daua Karajeanes, Ashley Prado, John Ruttan, Ivory Yang, Sean O'Brien, Vasu Sharma, and Kevin Zhu. 2025. Advancing uto-aztecan language technologies: A case study on the endangered comanche language. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 27–37.

Antonios Anastasopoulos, Christopher Cox, Graham Neubig, and Hilaria Cruz. 2020. Endangered languages meet modern nlp. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 39–45.

Steven Bird. 2020. Decolonising speech and language technology. In *28th International Conference on Computational Linguistics, COLING 2020*, pages 3504–3519. Association for Computational Linguistics (ACL).

Steven Bird. 2024. Must nlp be extractive? In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 14915–14929. Association for Computational Linguistics (ACL).

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Yi-Ping Chen. 1996. What are the functional orthographic units in chinese word recognition: The stroke or the stroke pattern? *The Quarterly Journal of Experimental Psychology: Section A*, 49(4):1024–1043.

Elena Suet-Ying Chiu. 2012. Heroines of jiangyong: Chinese narrative ballads in women's script.

Li Congrong. 2024. History, characteristics, and modern vitality of nüshu: A cultural anthropology perspective. *Anthropological Explorations of Gender, Identity, and Economics*, 85.

Fang Fan and Zhen Yong Lin. 1999. Online handwritten chinese character recognition system. In *Document Recognition and Retrieval VII*, volume 3967, pages 88–96. SPIE.

Imre Galambos. 2017. The chinese. *The Oxford Handbook of Classical Chinese Literature (1000 BCE-900 CE)*, page 31.

Google. 2020. Noto sans nushu. https://github.com/googlefonts/noto-cjk. Unicode font for Nüshu script.

Google Cloud. 2023. Google cloud vision api. https://cloud.google.com/vision. Accessed 2025.

Connie Suk-Han Ho and Peter Bryant. 1997. Learning to read chinese beyond the logographic phase. *Reading research quarterly*, 32(3):276–289.

Xihuan Hu. 2022. Authenticity issues in nüshu cultural heritage in china: Authentication, discourse, and identity-making. In *Cultures of Authenticity*, pages 43–61. Emerald Publishing Limited.

JaidedAI. 2020. Easyocr. https://github.com/JaidedAI/EasyOCR. Version 1.7.2.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. 2022. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*.

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

Xinrui Li. 2024. Nüshu: Scripting women's empowerment and cultural legacy in chinese society. *Journal of Theory and Practice of Social Science*, 4(03):21–23.

Yu Qin Lim and Azizi Bahauddin. 2025. From script to space: Integrating nüshu heritage into interior design narratives. *ARTEKS: Jurnal Teknik Arsitektur*, 10(1):53–62.

Fei-wen Liu. 2018. Practice and cultural politics of "women's script": nüshu as an endangered heritage in contemporary china. In *Women Writing Across Cultures*, pages 231–246. Routledge.

Xiaoyu Liu. 2024. The foreshadowing of women's tragic fate: an analysis of snow flower and the secret fan from the perspective of space theory. *International Journal of Linguistics, Literature and Translation*, 7(12):110–114.

Dylan Scott Low, Isaac Mcneill, and Michael Day. 2022. Endangered languages: A sociocognitive approach to language death, identity loss, and preservation in the age of artificial intelligence. *Sustainable Multilingualism*, 21(1):1–25.

BN Madhukar and R Narendra. 2013. Lanczos resampling for the digital processing of remotely sensed images. In *Proceedings of International Conference on VLSI, Communication, Advanced Devices, Signals & Systems and Networking (VCASAN-2013)*, pages 403–411. Springer.

Denise Schmandt-Besserat and Michael Erard. 2009. Origins and forms of writing. In *Handbook of research on writing*, pages 7–26. Routledge.

Annapurna Sharma and Dinesh Babu Jayagopi. 2021. Towards efficient unconstrained handwriting recognition using dilated temporal convolution network. *Expert Systems with Applications*, 164:114004.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633. IEEE.

Yuqian Sun, Yuying Tang, Ze Gao, Zhijun Pan, Chuyan Xu, Yurou Chen, Kejiang Qian, Zhigang Wang, Tristan Braud, Chang Hee Lee, and 1 others. 2023. Ai nüshu: An exploration of language emergence in sisterhood through the lens of computational linguistics. In *SIGGRAPH Asia 2023 Art Papers*, pages 1–7.

Yuying Tang, Yuqian Sun, Ze Gao, Zhijun Pan, Zhigang Wang, Tristan Braud, Chang Hee Lee, and Ali Asadipour. 2023. Ai nüshu (women's scripts)-an exploration of language emergence in sisterhood. In *ACM SIGGRAPH Asia 2023 Art Gallery*, pages 1–2.

Tencent Cloud. 2023. Tencent cloud optical character recognition (ocr). https://cloud.tencent.com/product/ocr. Accessed 2025.

The Unicode Consortium. 2017. The Unicode Standard, Version 10.0 – Nüshu. https://www.unicode.org/charts/PDF/U1B170.pdf. Accessed: 2025-05-18.

Jiangqing Wang and Rongbo Zhu. 2010. Handwritten nushu character recognition based on hidden markov model. *J. Comput.*, 5(5):663–670.

Xiaobo Wang. 2020. Nüshu, the unique female rhetoric in the chinese rhetorical tradition. In *The Routledge Handbook of Comparative World Rhetorics*, pages 297–305. Routledge.

Ivory Yang, Weicheng Ma, Carlos Guerrero Alvarez, William Dinauer, and Soroush Vosoughi. 2025a. What is it? towards a generalizable native american language identification system. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 105–111.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2025b. Nüshurescue: Reviving the endangered nüshu language with ai. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034.

Ivory Yang, Weicheng Ma, Chunhui Zhang, and Soroush Vosoughi. 2025c. Is it navajo? accurate language detection for endangered athabaskan languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 277–284.

Ivory Yang, Chunhui Zhang, Yuxin Wang, Zhongyu Ouyang, and Soroush Vosoughi. 2025d. Visibility as survival: Generalizing nlp for native alaskan language identification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6965–6979.

Haiwei Zhang. 2014. A review of stroke order in hanzi handwriting. *Language Learning in Higher Education*, 4(2):423–440.

Liming Zhao. 1992. *A Compendium of Chinese Nüshu (Zhongguo Nüshu Jicheng)*, original print edition, re-released as an ebook in 2019 edition. Tsinghua University Press, Beijing.

Liming Zhao. 1998. Nüshu: Chinese women's characters.

Wen Zuo and Metta Sirisuk. 2024. The jiangyong nüshu of china: Signifier and signified against the backdrop of indigenous intangible cultural heritage inheritance. *Journal of Roi Kaensarn Academi*, 9(2):189–206.

A Nüshu Language

A.1 Origins

Figure 3: Jiangyong county, Hunan, pointed out on a map.

Source: Jiangyong County (Wikipedia)

A.2 The Color Red

As evident throughout this work, the visual and stylistic themes of our paper draw inspiration from the cultural aesthetics of Nüshu. The script has associations with the color red, most notably vermilion (cinnabar) ink used on fans, paper, and red-dyed cloth for embroidered "third-day missives[4]".

Figure 4: Third-day missives.

Source: Atlas Obscura

[4]Third-day-missives are handwritten farewell letters written in Nüshu by women in Jiangyong County, China, as part of a traditional marriage ritual.

A.3 Nüshu Translations

Figure 5: A Nüshu poem with English translation: *"Beside a well, one does not thirst. Beside a sister, one does not despair."* This verse captures Nüshu as a manifestation of female expression and sisterhood.[A.3]

Source: PBS.org

B Community Collaboration

B.1 Art Creations and Exhibitions

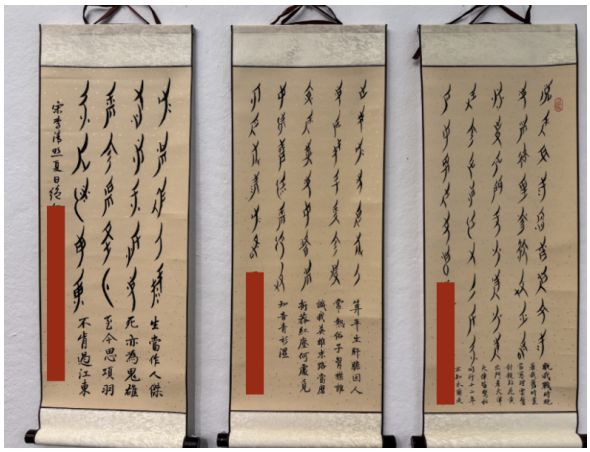


Figure 6: Three scrolls of Nüshu artwork created by our ancient Chinese calligrapher with 11 years of formal training.

Figure 7: Nüshu artifacts from Figure 6 featured at a recent art exhibition.

B.2 Interview Excerpt

> *"As a calligrapher with over 11 years of formal training, I have explored many traditional script styles. In the process of contributing to the Nüshu dataset, I was struck by the script's expressive beauty and historical depth. Nüshu characters exhibit a remarkable creativity and rhythm that sets them apart. Their strokes resemble those of Chinese Seal Script - elegant, varied, and full of artistic vitality, offering not only aesthetic appeal but also rich value for scholarly research."*

When working with calligraphers and fellow Nüshu scholars, we regularly conduct interviews to gather insights on the script's stylistic features, cultural significance, and practical considerations. These reflections not only enrich our understanding of Nüshu's artistic and historical context, but also inform the design and evaluation of our datasets. Feedback from practitioners ensures that our work remains grounded in the lived experience and embodied expertise of those most familiar with the script.

Specifically, our calligrapher noted that Nüshu's curvilinear, elongated strokes share stylistic similarities with seal script (Zhuanshu) in their rounded and decorative forms. Additionally, some simple, single-component Nüshu characters (e.g., the word for "inside") resemble regular script (Kaishu). These are descriptive observations, and not a claim of direct linguistic continuity.

C *NüshuStrokes* for Stroke Modeling

The *NüshuStrokes* dataset captures the sequential formation of all 397 Unicode Nüshu characters through frame-extracted handwriting videos. This temporal granularity enables a range of future applications.

Computationally, stroke order data can be used to model the dynamic characteristics of writing, such as directionality, stroke segmentation, and pen flow, which are critical for improving OCR and handwriting recognition models.

Pedagogically, the dataset can serve as a learning tool for teaching stroke order to new learners through real-time feedback and guided writing applications.

Culturally, it offers a visual archive of calligraphic practice, preserving not just what is written, but how it is written, a key step toward holistic revitalization of Nüshu.

D Revitalization and Benchmarking

D.1 Synthetic Data Considerations

Synthetic data is often used to address low-resource scenarios, but in endangered language contexts, it risks amplifying inauthentic patterns without native users to verify accuracy. Yang et al. (2025b) introduced a synthetic pipeline for Nüshu–Chinese translation, but acknowledged limitations in fluency and fidelity. As shown in Figure 8, synthetic outputs often reflected levels of inaccuracy. These challenges informed our decision to rely solely on expert-validated data for OCR, ensuring cultural and linguistic integrity.

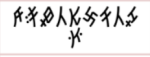
Examples	Chinese Input	Nüshu Translated Output
Correct	年时有人口万千人左右	
Incorrect	但这就是我当时最真实的感觉我不能逃避	
Not Translated (Length Mismatch)	调查员亦发现当时负责指示公务机的管制员并没有指示公务机下降至指定高度当时公务机应从尺下降至尺	—

Figure 8: Examples of synthetically translated Nüshu output from Yang et al. (2025b). Characters in pink are inaccurate, underscoring the difficulty of reliable generation without expert validation.

D.2 Toward Culturally Respectful Modeling

Language revitalization is not merely a technical challenge, it is a cultural, historical, and ethical one. Our work was conducted in close collaboration with Nüshu scholars and classically trained calligraphers, whose expertise informed every phase of the project, from data creation to model evaluation. In the absence of native validators, this collaborative knowledge transfer provides essential grounding for responsible modeling.