DNA: Dual-branch Network with Adaptation for Open-Set Online Handwriting Generation

Supplementary Material

In this supplementary material, we provide further details to enhance the understanding of our work. Specifically, we include:

1. **Detailed Notations:** Comprehensive definitions of all mathematical symbols and notations used throughout the manuscript.
2. **Qualitative Input Style Sample Size:** Additional qualitative comparisons demonstrating the impact of varying the number of input style samples on the generated handwriting.
3. **Qualitative Loss Combination:** Additional qualitative comparisons of generated handwriting resulting from different combinations of loss functions, providing visual evidence of the effect of each loss component.
4. **Qualitative Adaptive Content Branch:** Additional qualitative comparisons to examine the impact of local and global representations.
5. **Style–Content Trade-off:** Analysis in the trade-off relationship between style and content quality.
6. **Style-guided Synthesis:** User preference and qualitative results of style-guided synthesis.
7. **Ability and Limitation on Cross-linguistic:** Further experiments on cross-linguistic tasks demonstrating the extent to which DNA can perform in an open-set setting.
8. **Experiment on ProtoSnap benchmark:** Experiments on a small cuneiform sign dataset.
9. **Pre-processing of Decomposed Vectors:** Detailed explanations of the pre-processing of transferring each character into its structural and component vectors.
10. **Implementation Details of Evaluators:** Detailed explanations of the training strategy for the online content recognizer, offline content classification, and the style recognizer used in the main paper.

Additionally, for reference, the link to the Microsoft JhengHei font is https://learn.microsoft.com/zh-hk/typography/font-list/microsoft-jhenghei, and the Traditional Chinese online dataset is available at http://www.hcii-lab.net/data/scutcouch/.

1. Detailed Notations

To ensure facilitate readability throughout this paper, we provide a list of abbreviations and symbols in Tab. 1.

2. Qualitative Input Style Sample Size

Fig. 1 illustrates that by varying the content of style samples from the same writer, our model generates subtly different outputs. This variation aligns with real-life handwriting, where even when a person writes the same character multiple times, each instance will have slight differences. This flexibility allows our model to generate the natural diversity in handwriting styles, enhancing its realism.

3. Qualitative Loss Combination

Fig. 2 shows that overlooking $\mathcal{L}_{ct}$ causes the local and global encoders to lack clear guidance, leading to noise in the generated output. Applying $\mathcal{L}_{ct}$ helps define the character more clearly, making it closely resemble the target. However, stroke spacing remains inaccurate at both the character and writer levels. By incorporating the full setting, $\mathcal{L}_{sp}$ further refines the stroke spacing, resulting in a generated output that closely matches the target.

4. Qualitative Adaptive Content Branch

Fig. 3 illustrates that the local encoder effectively captures stroke-level details through component-based input, allowing the model to generate structurally correct characters even without seeing full images. In contrast, the global encoder, guided by actual character images, better preserves the overall character shape, but may generate less accurate stroke details. By combining both, the full model integrates fine-grained stroke control with global structural guidance, resulting in characters that are not only accurate and recognizable but also closely aligned with the target style.

5. Style–Content Trade-off

Style and content performance are inherently interdependent: mimicking a "scrawled" style improves handwriting imitation but often harms recognizability, and vice versa. As shown in Fig. 4 (blue squares), in examples (1, 2, and 6, from left to right), SDT closely mirrors the global style of the ground truth (GT), producing strokes that approximate the GT well and achieving a higher style score. However, this comes at the cost of structural precision, resulting in incorrect content. In contrast, DNA achieves a better balance between style and content, as highlighted in the (red squares), which facilitates accurate character generation. This improvement stems from the design of the content input: SDT relies solely on global character im-

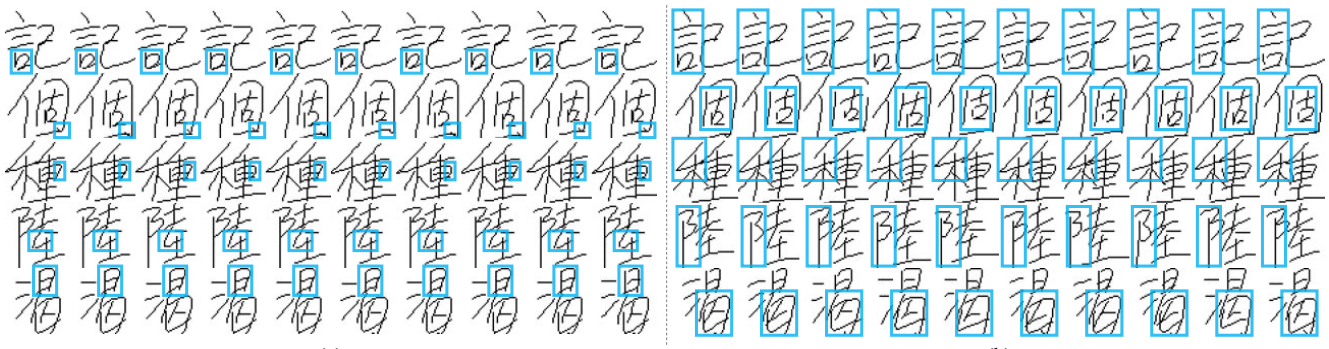
Notation / Abbreviation	Description
Overall Task and Settings	
OOHG	Open-set Online Handwriting Generation
DNA	Dual-branch Network with Adaptation
SWSC	Seen-Writer-Seen-Characters
UWSC	Unseen-Writer-Seen-Characters
UWUC	Unseen-Writer-Unseen-Characters
Datasets, Inputs, and Outputs	
$X_s = \{x_w^s\}_{w=1}^{N_w}$	Style dataset of N_w writers.
$x_w^s = \{I_0^w, \ldots, I_{N_s-1}^w\}$	The set of N_s style images of the w-th writer.
$X_c = \{(\mathbf{v}_i^c, I_i^c)\}_{i=1}^{N_c}$	Content dataset of N_c characters.
$\mathbf{v}_i^c = \{\mathbf{v}_i^{struct}, \mathbf{v}_i^{compo}\}$	Decomposition vectors for the i-th character (structure and components).
$I_i^c \in \mathbb{R}^{3\times H\times W}$	Printed character image for the i-th character.
$Y_{w,i} = \{p_j^{w,i}\}_{j=1}^{N_p}$	Ground-truth online handwriting (point set) for writer w, character i.
$p_j^{w,i} = \{\Delta x_j, \Delta y_j, m_j^1, m_j^2, m_j^3\}$	Each point's offsets $(\Delta x_j, \Delta y_j)$ and pen states m_j^1, m_j^2, m_j^3.
Style Branch and Content Branch	
$E_s(\cdot), h_s(\cdot), h_g(\cdot)$	Style encoder, style head, and glyph head of the style branch.
$Z_w^s = E_s(x_w^s)$	Extracted style features from writer w's style images.
$S_w^s = h_s(Z_w^s)$	Writer-specific style (S_w^s) features.
$G_w^s = h_g(Z_w^s)$	Glyph-specific style (G_w^s) features.
$E_c^l(\cdot)$	Local (structural) encoder.
$E_c^g(\cdot)$	Global (texture) content encoder.
$Z_i^{lc} = E_c^l(\mathbf{v}_i^c)$	Local representation (structure + component) for character i.
$Z_i^{gc} = E_c^g(I_i^c)$	Global texture representation for character i.
Z_i^c	Fused content feature after cross-attention between Z_i^{lc} and Z_i^{gc}.
Decoder and Predictions	
$\hat{p}_t^{w,i}$	Predicted point at step t for writer w, character i.
$\hat{Y}_{w,i} = \{\hat{p}_t^{w,i}\}_{t=0}^{N_p}$	Generated online handwriting sequence.
Loss Functions	
$\mathcal{L}_{xy}, \mathcal{L}_{state}$	Negative log-likelihood and cross-entropy losses for point prediction.
$\mathcal{L}_{sty}, \mathcal{L}_{gly}$	Style and glyph contrastive losses.
$\mathcal{L}_{ch}, \mathcal{L}_{de}$	Character classification and decomposition losses (part of $\mathcal{L}_{ct}$).
$\mathcal{L}_{ct} = \mathcal{L}_{ch} + \mathcal{L}_{de}$	Content loss.
$\mathcal{L}_{sp}$	Spacing loss.
$\mathcal{L}_{base} = 2\mathcal{L}_{state} + \mathcal{L}_{xy} + \mathcal{L}_{sty} + \mathcal{L}_{gly}$	Baseline loss.

Table 1. Notations and abbreviations

ages, which may overlook fine-grained structural details, whereas DNA integrates both local component information and global content images, leading to better preservation of stroke details and local structure — ultimately producing results that are more faithful to both the local style and content of the GT. Nevertheless, this can result in a slight reduction in style score.

6. Style-guided Synthesis

In Tab. 2, we conduct a user study involving 50 volunteers for voting on which generated samples are most similar to the ground truths. We received 47% and 53% user preference in the UWSC and UWUC settings, respectively. Additionally, we provide qualitative results for style-guided synthesis in Fig. 5. When the style references come from different writers, our model successfully generates other characters that faithfully reflect each writer's unique handwriting



(a) (b)

Figure 1. (a) Generated results using 15 style samples randomly selected from 16 style samples. (b) Generated results using 15 style samples randomly selected from 30 style samples. The blue squares mark the subtle differences between each inference sample.

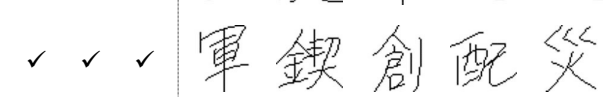
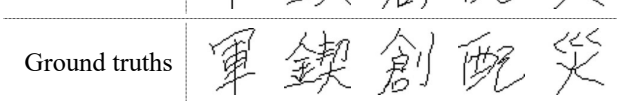
Settings			UWSC					UWUC				
$\mathcal{L}_{base}$	$\mathcal{L}_{ct}$	$\mathcal{L}_{sp}$	軍	鍥	創	配	災	勻	昕	疣	偯	驍
✓												
✓	✓											
✓		✓										
✓	✓	✓										
Ground truths												

Figure 2. Qualitative of DNA with different loss combinations under the UWSC and UWUC settings.

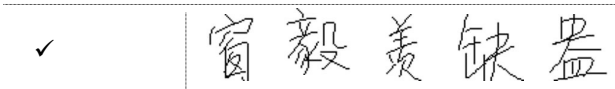
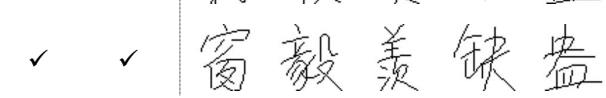
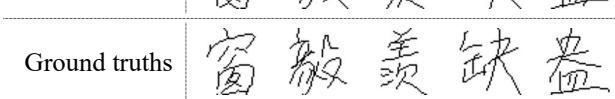
Content branch		UWSC					UWUC				
Local	Global	窗	毅	羨	缺	盎	莧	訾	鎰	癟	鍔
✓											
	✓										
✓	✓										
Ground truths											

Figure 3. Qualitative of DNA with different combinations on the adaptive content branch under the UWSC and UWUC settings.

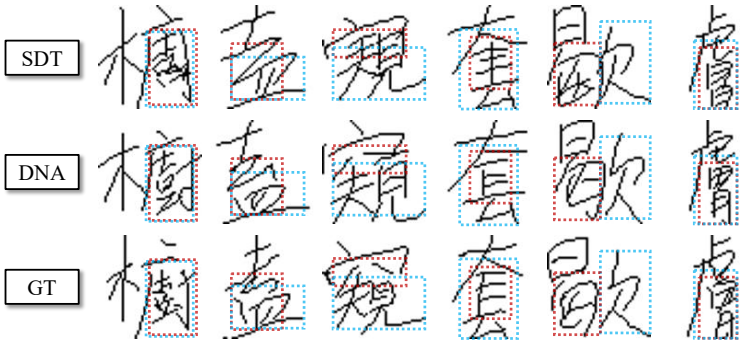


Figure 4. Qualitative results highlighting the trade-off between style and content scores. DNA seeks a balanced outcome between style and content, whereas SDT tends to sacrifice content accuracy to improve the style score. The red and blue boxes indicate regions of interest for evaluating content correctness and style similarity to the ground truth (GT), respectively.

Method	User Study (%)	
	UWSC	UWUC
FontDiffuser	24.78	18.23
WriteLikeYou	2.61	1.47
SDT	25.22	27.06
DNA	**47.39**	**53.24**

Table 2. Handwriting generation results evaluated by user preference, showing the percentage of volunteers who judged the generated samples as more similar to the ground truth.

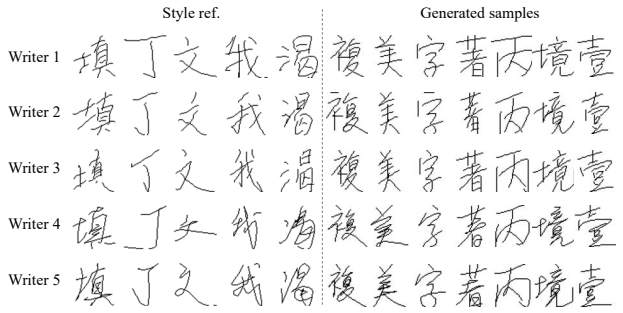


Figure 5. Qualitative results of style-guided synthesis. The left side displays style reference inputs, while the right side shows the generated samples in the writers' style.

style. Notably, for the fourth and fifth writers, whose references are drawn from another dataset characterized by more cursive handwriting, our method still captures the stylistic traits and produces characters that closely resemble the target writing style.

7. Ability and Limitation on Cross-linguistic

Cross-linguistic handwriting generation is challenging because character components differ across languages. Our method synthesizes unseen characters by recombining known components; however, generating characters composed entirely of unseen components remains difficult. In

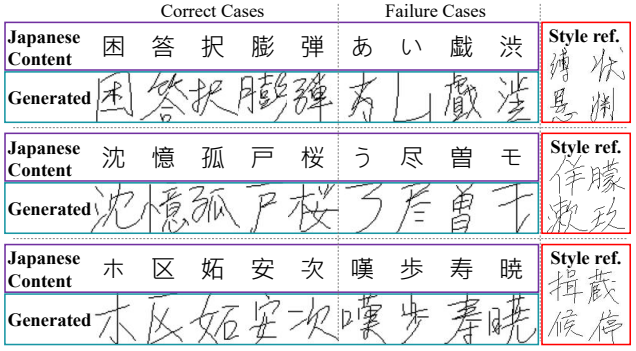


Figure 6. Generalization of cross-language, where DNA is trained on Chinese characters and directly generates Japanese characters. The rightmost column shows the style reference.

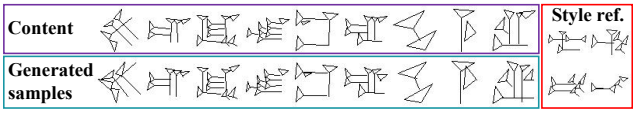


Figure 7. Generalization of the ProtoSnap benchmark [4], where DNA is well-trained on Chinese characters and fine-tuned on cuneiform signs. The rightmost column shows the style reference.

Fig. 6, we evaluate our model trained on traditional Chinese for generating Japanese characters. As shown on the left, characters sharing components with Chinese remain recognizable, whereas those composed of entirely novel components result in failed generations.

8. Experiment on ProtoSnap benchmark

We fine-tune DNA, originally well-trained on Chinese, using the ProtoSnap benchmark [4], which provides prototype alignment for cuneiform signs. For fine-tuning and evaluation, we adopt the Assurbanipal subset of ProtoSnap, which contains 161 valid characters, with 131 used for training and 30 reserved for testing. In the case of cuneiform signs, the decompositions are straightforward since their components are explicitly reflected in the trajectory data, mak-

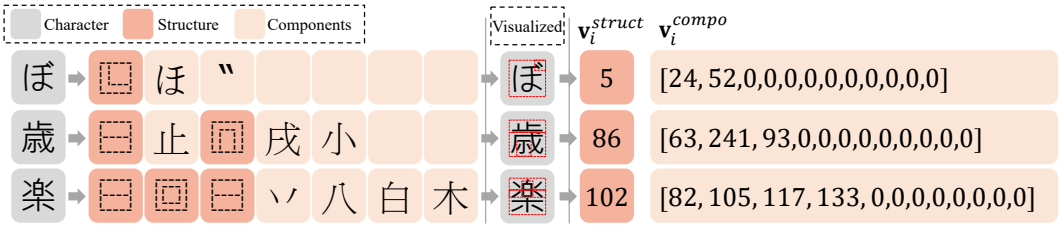


Figure 8. Examples of decomposition of Japanese characters and their corresponding encoding vectors, $\mathbf{v}_i^{struct}$ and $\mathbf{v}_i^{compo}$.

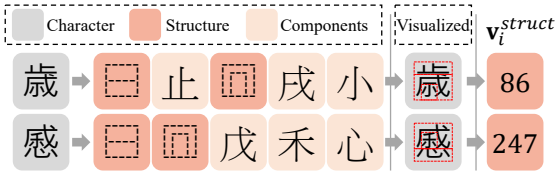


Figure 9. Examples of different structural vectors from similar structure decompositions.

ing it possible to identify both the components and structural organization of each character. This property makes cuneiform signs particularly suitable for DNA to naturally leverage component-structure alignments to enhance generation quality. The generated samples are shown in Fig. 7.

9. Pre-processing of Decomposed Vectors

Following the IDS (Ideographic Description Sequence) rules, both Chinese and Japanese characters can be decomposed into structures and components, as illustrated in Fig. 3 of the main paper and Fig. 8, respectively. We separate them to construct a structural vector and a component vector.

Structural vector. This vector is determined by all structural elements (dark orange) of a character and their relative positions. For example, as shown in Fig. 9, the structure ⿰ in the second position differs from ⿰ in the third position, and thus their structural indices are assigned differently to reflect their distinct roles in the character composition.

Component vector. Each component element corresponds to a unique component index, and the component vector is constructed by sequentially arranging these indices. Zero-padding is applied where necessary.

10. Implementation Details of Evaluators

We split the UWSC dataset into 80% for training and 20% for testing. Similarly, the UWUC dataset is divided using the same 8:2 ratio. Content evaluators are then trained on each dataset setting, and the trained evaluators are used to assess each set separately.

1. **Dynamic Time Warping (DTW)** [1, 2]: We employ DTW as an elastic matching method to align two sequences, calculating the distance between the generated and actual characters. This alignment-based approach allows for a robust comparison of sequence similarity between generated and real characters.
2. **Content Score (CS)** [6]: To assess structural accuracy, we utilize a content recognizer trained specifically on the dataset's training set to evaluate how accurately the generated character structures match the intended characters. The recognizer uses the Adam optimizer with a learning rate of 0.001 and a batch size of 128. Additionally, for offline methods, we use images as input for training, whereas for online methods, we use a coordinated sequence of characters for training.
3. **Fréchet Inception Distance (FID)**: To quantitatively assess the stylistic similarity between generated and real handwriting, we use FID. This metric compares the distributions of high-level visual features, capturing both the realism and stylistic coherence of the generated handwriting.
4. **Geometry Score (GS)** [3]: To complement the FID-based evaluation, we use GS to assess the structural of generated handwriting. This metric compares the topological properties of the generated and real data manifolds, revealing how well the model captures the underlying geometry of the handwriting style.
5. **Character Error Rate (CER)**: To assess the benefit of generated data for downstream handwriting recognition, we trained a Handwritten Text Recognition (HTR) model [5] with a combination of real and generated samples. The real set comes from the **SWSC** setting and includes all 4807 characters from each of 55 writers. For the generated data, two settings were used: in **UWSC**, we synthesized 4807 characters for each of 10 writers (48070 samples in total) and evaluated the model on the **UWSC** ground truth; in **UWUC**, we generated 594 unseen characters for each of 10 unseen writers (5940 samples) and tested on the corresponding ground-truth set.

References

[1] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pages 359–370, 1994. 5

[2] Zhounan Chen, Daihui Yang, Jinglin Liang, Xinwu Liu, Yuyi Wang, Zhenghua Peng, and Shuangping Huang. Complex handwriting trajectory recovery: Evaluation metrics and algorithm. In *ACCV*, pages 1060–1076, 2022. 5

[3] Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. In *International conference on machine learning*, pages 2621–2629. PMLR, 2018. 5

[4] Rachel Mikulinsky, Morris Alper, Shai Gordin, Enrique JimÃŠnez, Yoram Cohen, and Hadar Averbuch-Elor. Protosnap: Prototype alignment for cuneiform signs. In *ICLR*, 2025. 4

[5] George Retsinas, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. Best practices for a handwritten text recognition system. In *International Workshop on Document Analysis Systems*, pages 247–259. Springer, 2022. 5

[6] Bocheng Zhao, Jianhua Tao, Minghao Yang, Zhengkun Tian, Cunhang Fan, and Ye Bai. Deep imitator: Handwriting calligraphy imitation via deep attention networks. *PR*, 104:107080, 2020. 5