GLYPH NORMALIZATION WITH DISCRETE GENERATIVE TRANSPORT FOR LOW-RESOURCE TANGUT CURSIVE SCRIPT

Anonymous authors
Paper under double-blind review

ABSTRACT

Tangut cursive glyphs exhibit stroke connections, omissions, compression, and structural deformation, making direct recognition difficult in low-resource settings. We formulate cursive processing as *glyph normalization*: transforming a cursive glyph into its canonical printed form while preserving character structure. We propose the Source Conditioned Generative Transport Model (SC-GTM), which represents cursive and printed glyphs in a shared discrete visual-token space and performs iterative many-to-many spatial transport followed by bounded local token rewriting. Transport reorganizes visible local evidence, whereas rewriting accounts for morphological changes and partially missing structures. Because paired data specify only the source and target endpoints, we regularize the intermediate dynamics with path relative entropy to a fixed reference process, together with paired endpoint supervision. On both unseen-character and seen-character/unseen-instance protocols, SC-GTM obtains favorable structural consistency, especially on normalized skeleton Chamfer distance (NSCD) and foreground IoU. These results support explicit spatial reorganization and local rewriting as a useful formulation for low-resource Tangut cursive normalization.

1 INTRODUCTION

Tangut documents are important sources for studying the history, language, and social life of the Western Xia. Cursive Tangut is particularly difficult to interpret because stroke connections, omissions, compression, and deformation can substantially change local structure and spatial layout. Since reliably transcribed cursive data are scarce, a practical model should learn reusable transformation regularities rather than memorize character classes.

Cursive-to-print conversion is therefore not simple style transfer. The canonical target may require both *reorganizing* evidence that remains visible in the cursive input and *recovering* local structures that are strongly altered or omitted. Fixed one-to-one correspondences between source and target strokes, components, or positions are generally unavailable, and intermediate restoration trajectories are not annotated.

We address this problem with a Source Conditioned Generative Transport Model (SC-GTM). Cursive and printed glyphs are encoded into a shared discrete visual-token space. At each restoration round, a source-conditioned transport module redistributes local probability states under capacity constraints, after which a locally conditioned rewriting module changes token categories with a bounded residual update. We further regularize the complete transport–rewrite trajectory by its path relative entropy to a fixed reference process, while supervising only the paired endpoint.

Our contributions are:

- We formulate low-resource Tangut cursive normalization as dynamic structural restoration with unknown local correspondence, separating spatial reorganization from local category change.
- We introduce a many-to-many transport–rewrite process in a shared discrete token space, without stroke, component, or intermediate-trajectory annotations.

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.

- We combine paired endpoint supervision with path-space regularization inspired by Schrödinger bridges, enabling multistep restoration while keeping the reference dynamics fixed.

2 Related Work

Tangut and low-resource recognition. Prior Tangut work mainly addresses database construction and character classification (Ma et al., 2022; Ma and Ma, 2023), while cursive-script datasets focus on recognizing character or style labels (Liang et al., 2020; Shan et al., 2024; Zhao et al., 2025). Low-resource handwriting recognition has benefited from pretraining and data-efficient learning (Li et al., 2023b; Vogler et al., 2022; Souibgui et al., 2022; Li et al., 2025). In contrast, we predict a canonical printed glyph and explicitly model the structural transformation from incomplete cursive evidence.

Glyph generation and restoration. Image translation (Isola et al., 2017; Zhu et al., 2017; Cao et al., 2023) and few-shot font generation (Park et al., 2021a;b; Kong et al., 2022; Xie et al., 2021; Tang et al., 2022; Yao et al., 2024) learn cross-style or local structural regularities, but typically assume clean content glyphs or target font generation rather than canonical restoration from cursive input. Diffusion and bridge models provide multistep alternatives (Saharia et al., 2022; Yang et al., 2024; Liu et al., 2023; Li et al., 2023a). Our formulation instead separates spatial redistribution from local category change in a discrete position–token state space. It uses Sinkhorn projection (Cuturi, 2013) for capacity-constrained transport and path relative entropy inspired by Schrödinger bridges (Léonard, 2014; Shi et al., 2023; Ksenofontov and Korotin, 2025); it does not claim to solve a standard Schrödinger bridge.

3 Method

3.1 Task and discrete probabilistic state

Let

$$\mathcal{D} = \{(x_n^c, x_n^r)\}_{n=1}^M \tag{1}$$

contain a cursive image x^c and its canonical printed image x^r. The training pairs provide no local correspondences or intermediate trajectories. A shared vector-quantized encoder E and decoder D map both domains to the same $H \times W$ token grid,

$$X = E(x^c), \qquad Y = E(x^r), \qquad X, Y \in \{1, \dots, K\}^{H \times W}, \tag{2}$$

where K is the codebook size and $N = HW$. The tokenizer is trained for reconstruction and then frozen while learning the dynamics.

Rather than selecting one token at every intermediate step, the model maintains a position-wise probability map $p_t \in [0,1]^{N \times K}$. We initialize

$$p_0(i,a) = \mathbf{1}\{a = X_i\}, \qquad \mu_t(i,a) = \frac{p_t(i,a)}{N}, \qquad \sum_a p_t(i,a) = 1. \tag{3}$$

The full K-dimensional state is retained throughout the dynamics, and the original source grid X remains available to the transport controller.

3.2 Capacity-constrained spatial transport

At round t, a global controller reads (p_t, X, t) and produces transport scores $s_t^m(i,a,j)$ on a fixed symmetric multiscale graph with self-loops. Let $F_t(i,a,j)$ be the mass moved from local state (i,a) to position j. We impose

$$\sum_j F_t(i,a,j) = \mu_t(i,a), \qquad \sum_{i,a} F_t(i,a,j) = \frac{1}{N}. \tag{4}$$

The first constraint preserves outgoing mass, while the second gives each destination position equal capacity. Hence one source state may split across destinations and several sources may merge at one position.

Let $A^0(i,j)$ be a fixed reference spatial kernel that favors short movements. We form

$$W_t(i,a,j) = \mu_t(i,a)A^0(i,j)\exp\left(\frac{s_t^m(i,a,j)}{\tau_m}\right), \tag{5}$$

$$F_t = \arg\min_{F\in\mathcal{U}_t} D_{\mathrm{gen}}(F\|W_t), \tag{6}$$

where $\mathcal{U}_t$ is the feasible set in Eq. (4). The projection is computed by Sinkhorn row–column scaling; feasibility and the scaling form are given in Appendix B.1. The transported distribution is

$$\widetilde{\mu}_t(j,a) = \sum_i F_t(i,a,j), \qquad \widetilde{p}_t(j,a) = N\widetilde{\mu}_t(j,a). \tag{7}$$

3.3 Bounded local rewriting

Transport changes where local states are placed but cannot by itself create a token category that is absent at the required location. We therefore apply category rewriting after transport. To prevent rewriting from becoming an independent global generator, its controller reads only the transported 3×3 neighborhood around position j and produces a row-stochastic matrix $C_t(j)$.

The actual rewriting kernel is a bounded residual mixture

$$B_t(j) = (1-\rho)I + \rho C_t(j), \qquad 0<\rho<\tfrac{1}{2}, \tag{8}$$

with update

$$p_{t+1}(j,\cdot) = (1-\rho)\widetilde{p}_t(j,\cdot) + \rho\,\widetilde{p}_t(j,\cdot)C_t(j). \tag{9}$$

Thus each round preserves at least a $(1-\rho)$ fraction of the arrived category mass while still permitting transitions between any token categories. For a fixed rewriting kernel, distinct inputs satisfy $\|uB - vB\|_1 \geq (1-2\rho)\|u-v\|_1$; the derivation and its scope are provided in Appendix B.2. Alternating Eqs. (7) and (9) enables progressive spatial and morphological correction.

3.4 Path regularization, endpoint supervision, and inference

Endpoint pairs do not uniquely determine the internal restoration path. We therefore regularize the dynamics against a fixed reference process that alternates A^0 with

$$B^0(a,b) = (1-\eta)\mathbf{1}\{a=b\} + \frac{\eta}{K}, \qquad 0<\eta<\rho. \tag{10}$$

For a source grid X, the learned transition sequence defines a local-state path distribution P_θ^X, and the reference transitions define R^X. Its path relative entropy decomposes exactly into edge-flow terms:

$$\mathcal{L}_{\mathrm{path}} = \mathrm{KL}(P_\theta^X\|R^X) \tag{11}$$

$$= \sum_{t=0}^{T-1}\left[\sum_{i,a,j} F_t(i,a,j)\log\frac{F_t(i,a,j)}{\mu_t(i,a)A^0(i,j)} + \sum_{j,a,b}\widetilde{\mu}_t(j,a)B_t(j,a,b)\log\frac{B_t(j,a,b)}{B^0(a,b)}\right]. \tag{12}$$

The target is used only through soft endpoint supervision,

$$\mathcal{L}_{\mathrm{end}} = -\frac{1}{N}\sum_i \log p_T(i,Y_i), \qquad \min_\theta\ \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\mathcal{L}_{\mathrm{end}} + \lambda\mathcal{L}_{\mathrm{path}}\right]. \tag{13}$$

The decomposition and finite-step implications are detailed in Appendices B.3–B.4. At inference, $\hat{Z}_i = \arg\max_a p_T(i,a)$ and the normalized printed image is $\hat{x}^p = D(\hat{Z})$. No target image or path-KL computation is required. The dynamic provenance/correspondence construction used for analysis is moved to Appendix B.5.

Table 1: Main structural comparison. Mean ± standard deviation. Best is **bold**; second best is underlined.

	Unseen Character		Seen Char. / Unseen Inst.	
Method	NSCD ↓	IoU ↑	NSCD ↓	IoU ↑
Pix2Pix	0.0182	0.2486	0.0212	0.2107
CycleGAN	0.03018	0.1228	0.03608	0.1059
Pix2Pix-Turbo	0.03677	0.2487	0.04053	0.2216
Palette	0.02592	0.1504	0.02640	0.1550
VQ-Diffusion	0.02137	0.1975	0.02349	0.1854
Flow Matching	0.03388	0.2561	0.03959	0.2213
DDBM	0.03647	0.2864	0.04632	**0.2420**
I^2SB	0.08465	0.2052	0.09654	0.1847
CSBM	0.05264	0.1545	0.06924	0.1246
SC-GTM	**0.01602**	**0.2919**	**0.02027**	0.2315

4 EXPERIMENTS

4.1 SETUP

The paired dataset contains 1,048 Tangut character classes, split into 890 training, 79 validation, and 79 test classes, with all glyph images normalized to 128×128 pixels. We evaluate two protocols: **Unseen Character**, where none of the 79 test character classes occurs in training, and **Seen Character / Unseen Instance**, where the character class is observed during training but the evaluated cursive realization is held out. The tokenizer is trained only on the training and validation images; unseen-character test prints and held-out cursive instances are excluded from tokenizer and SC-GTM training. Full protocol details are in Appendix A.

We report foreground IoU, normalized skeleton Chamfer distance (NSCD; lower is better), SSIM, and LPIPS. The main text emphasizes the two structure-oriented metrics; the complete four-metric table is in Appendix A.3. Baselines cover paired/unpaired image translation, diffusion/flow models, and bridge-based models: Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017), Pix2Pix-Turbo (Parmar et al., 2024), Palette (Saharia et al., 2022), VQ-Diffusion (Gu et al., 2022), Flow Matching (Lipman et al., 2023), DDBM (Zhou et al., 2024), I^2SB (Liu et al., 2023), and CSBM (Ksenofontov and Korotin, 2025).

4.2 MAIN RESULTS

Table 1 shows that SC-GTM has its clearest advantage on structural measures. On Unseen Character, it obtains the lowest NSCD (0.01602) and highest foreground IoU (0.2919), indicating that the generated skeleton locations and foreground support remain close to the canonical targets even for classes absent from training. On Seen Character / Unseen Instance, SC-GTM again has the lowest NSCD (0.02027) and the second-highest IoU (0.2315), behind DDBM (0.2420).

The complete results show a different ordering on appearance/perceptual metrics: Pix2Pix attains the highest SSIM in both protocols, while Flow Matching attains the lowest LPIPS. Thus, global image similarity and local structural fidelity are not interchangeable for this task. We therefore report all four metrics but use NSCD and foreground IoU as the primary evidence for structural normalization. Full numerical results and qualitative comparisons are provided in Appendices A.3 and A.4.

4.3 ABLATION

Replacing learned transport with identity transport degrades both NSCD and IoU, showing that local rewriting alone is insufficient for the observed spatial changes. Removing path regularization also degrades all four reported metrics, although less strongly. These ablations support complementary roles for explicit spatial reorganization and path-level regularization.

Table 2: Ablation on the Unseen Character test set. Mean ± standard deviation.

Variant	NSCD ↓	IoU ↑	SSIM ↑	LPIPS ↓
w/o Transport	0.0218±0.0031	0.2594±0.0082	0.5216±0.0068	0.2698±0.0041
w/o Path Reg.	0.0187±0.0027	0.2768±0.0071	0.5309±0.0062	0.2561±0.0037
SC-GTM	**0.01602**±0.00226	**0.2919**±0.0065	**0.5377**±0.0059	**0.2475**±0.0033

5 Conclusion

We formulated low-resource Tangut cursive normalization as multistep structural restoration and introduced SC-GTM, which alternates capacity-constrained many-to-many transport with bounded local token rewriting in a shared discrete representation. Paired endpoint supervision specifies the desired printed glyph, while path relative entropy regularizes otherwise underdetermined intermediate dynamics. Across unseen-character and unseen-instance evaluations, the method is strongest on skeleton and foreground-structure measures, whereas some baselines remain better on SSIM or LPIPS. The method is limited by tokenizer quality, the absence of a sufficiently reliable Tangut recognizer for direct identity evaluation, and severe cases where the cursive input omits too much local evidence. The learned transport should be interpreted as probabilistic provenance within the model rather than literal stroke genealogy.

AI Use Statement

During this research, generative AI tools were used to assist with methodological discussion, verification of mathematical formulations, initial drafting, and language polishing. All technical content, derivations, references, experimental protocols, results, and conclusions were reviewed and verified by the authors, who take full responsibility for the manuscript.

References

Y. Cao, J. Ma, C. Hao, and Q. Yan. Tangut character image generation based on cycle-consistent adversarial networks. *Journal of Intelligent & Fuzzy Systems*, 44(4):6341–6358, 2023. doi: 10.3233/JIFS-221892.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10696–10706, 2022.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

Y. Kong, C. Luo, W. Ma, Q. Zhu, S. Zhu, N. Yuan, and L. Jin. Look closer to supervise better: One-shot font generation via component-based discriminator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13482–13491, 2022.

G. Ksenofontov and A. Korotin. Categorical schrödinger bridge matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 31727–31751, 2025.

C. Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34(4):1533–1574, 2014. doi: 10.3934/dcds.2014.34.1533.

B. Li, K. Xue, B. Liu, and Y.-K. Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1952–1961, 2023a.

M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102, 2023b. doi: 10.1609/aaai.v37i11.26538.

Y. Li, D. Chen, T. Tang, and X. Shen. Htr-vt: Handwritten text recognition with vision transformer. *Pattern Recognition*, 158:110967, 2025. doi: 10.1016/j.patcog.2024.110967.

J. Liang, W.-H. Liao, and Y.-C. Wu. Toward automatic recognition of cursive chinese calligraphy: An open dataset for cursive chinese calligraphy text. In *Proceedings of the 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–5, 2020. doi: 10.1109/IMCOM48794.2020.9001777.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

G.-H. Liu, A. Vahdat, D.-A. Huang, E. Theodorou, W. Nie, and A. Anandkumar. I2sb: Image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22042–22062, 2023.

J. Ma, Y. Cao, Z. Ma, L. Wei, and C. Hao. End-to-end tangut character database building and recognition method. *IET Image Processing*, 16(8):2087–2100, 2022. doi: 10.1049/ipr2.12471.

Z. Ma and J. Ma. Ttn-fcn: A tangut character classification framework by tree tensor network and fully connected neural network. *IET Image Processing*, 17(13):3815–3829, 2023. doi: 10.1049/ipr2.12899.

S. Park, S. Chun, J. Cha, B. Lee, and H. Shim. Few-shot font generation with localized style representations and factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2393–2402, 2021a. doi: 10.1609/aaai.v35i3.16340.

S. Park, S. Chun, J. Cha, B. Lee, and H. Shim. Multiple heads are better than one: Few-shot font generation with multiple localized experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13900–13909, 2021b.

G. Parmar, T. Park, S. Narasimhan, and J.-Y. Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.

C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. doi: 10.1145/3528233.3530757.

X. Shan, K. Zhang, L. Shen, and B. Wang. The wushu database for cursive script character and style recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2024. doi: 10.1109/ICMEW63481.2024.10645415.

Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet. Diffusion schrödinger bridge matching. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

M. A. Souibgui, A. Fornés, Y. Kessentini, and B. Megyesi. Few shots are all you need: A progressive learning approach for low resource handwritten text recognition. *Pattern Recognition Letters*, 160: 43–49, 2022. doi: 10.1016/j.patrec.2022.06.003.

L. Tang, Y. Cai, J. Liu, Z. Hong, M. Gong, M. Fan, J. Han, J. Liu, E. Ding, and J. Wang. Few-shot font generation by learning fine-grained local styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7895–7904, 2022.

N. Vogler, J. Allen, M. Miller, and T. Berg-Kirkpatrick. Lacuna reconstruction: Self-supervised pre-training for low-resource historical document transcription. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 206–216, 2022. doi: 10.18653/v1/2022.findings-naacl.15.

Y. Xie, X. Chen, L. Sun, and Y. Lu. Dg-font: Deformable generative networks for unsupervised font generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5130–5140, 2021.

Z. Yang, D. Peng, Y. Kong, Y. Zhang, C. Yao, and L. Jin. Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6603–6611, 2024. doi: 10.1609/aaai.v38i7.28482.

M. Yao, Y. Zhang, X. Lin, X. Li, and W. Zuo. Vq-font: Few-shot font generation with structure-aware enhancement and quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16407–16415, 2024. doi: 10.1609/aaai.v38i15.29577.

Y. Zhao, Y. Zhang, and L. Jin. Mccd: A multi-attribute chinese calligraphy character dataset annotated with script styles, dynasties, and calligraphers. *arXiv preprint arXiv:2507.06948*, 2025.

L. Zhou, A. Lou, S. Khanna, and S. Ermon. Denoising diffusion bridge models. In *International Conference on Learning Representations (ICLR)*, 2024.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. doi: 10.1109/ICCV.2017.244.

A Experimental Details

A.1 Dataset and protocols

To our knowledge, no public dataset directly provides paired Tangut cursive-to-print normalization at the required scale. We therefore collected and curated paired cursive/printed character images from facsimiles of printed sources and publicly available online materials. Each character was cropped and normalized to 128×128 pixels. The dataset contains 1,048 character classes, split into 890/79/79 training/validation/test classes.

Seen Character / Unseen Instance. We select 79 training-side character classes with at least two cursive realizations. Training uses the designated cursive–printed pair; one additional cursive realization per selected class is excluded from all model training and used for evaluation. The character identity is therefore observed during training, but the evaluated cursive instance is not.

Unseen Character. None of the 79 test classes appears in model training. The model must restore these new character structures using regularities learned from other classes. The tokenizer is trained only on training/validation images: neither cursive nor printed images of unseen test classes participate in tokenizer training, and held-out cursive instances from the seen-character protocol are also excluded from tokenizer and SC-GTM training.

Because the experimental character set is not covered by a sufficiently reliable Tangut recognizer, we avoid classifier-based recognition accuracy. Foreground IoU measures overlap of binarized glyph regions. NSCD skeletonizes prediction and target, computes bidirectional Chamfer distance, and normalizes by the image diagonal $128\sqrt{2}$. SSIM measures image-level structural similarity, and LPIPS measures perceptual difference. Metrics are computed per image and then averaged; repeated independent runs are reported as mean $\pm$ standard deviation.

A.2 Implementation and baselines

All methods use the same splits and 128×128 input resolution, with public implementations or settings recommended by their original papers, plus the interface changes required to use cursive images as inputs and printed images as supervised targets. Stochastic methods are evaluated on the same test samples with fixed random seeds. Detailed network and hyperparameter settings should remain those used in the original experimental code.

The compared methods are Pix2Pix (Isola et al., 2017), Pix2Pix-Turbo (Parmar et al., 2024), CycleGAN (Zhu et al., 2017), Palette (Saharia et al., 2022), VQ-Diffusion (Gu et al., 2022), Flow Matching (Lipman et al., 2023), DDBM (Zhou et al., 2024), I^2SB (Liu et al., 2023), and CSBM (Ksenofontov and Korotin, 2025).

A.3 Full quantitative comparison

A.4 Qualitative comparison

The complete visual comparisons are moved out of the main paper because they occupy nearly two full pages. Under Unseen Character, several methods produce printed-looking outputs while retaining local structural errors; SC-GTM often preserves the dominant skeleton and foreground layout more closely, consistent with the structural metrics. Under Seen Character / Unseen Instance, most paired methods generate recognizable forms, but local shifts and missing details remain for difficult cursive realizations.

B Supplementary Derivations and Analysis

B.1 Reference transport, capacity projection, and state validity

Let allowed spatial connections form an undirected graph with self-loops and symmetric positive weights $w_{ij} = w_{ji}$ on non-diagonal allowed edges. With

$$d_{\max} = \max_i \sum_{j \neq i} w_{ij}, \tag{14}$$

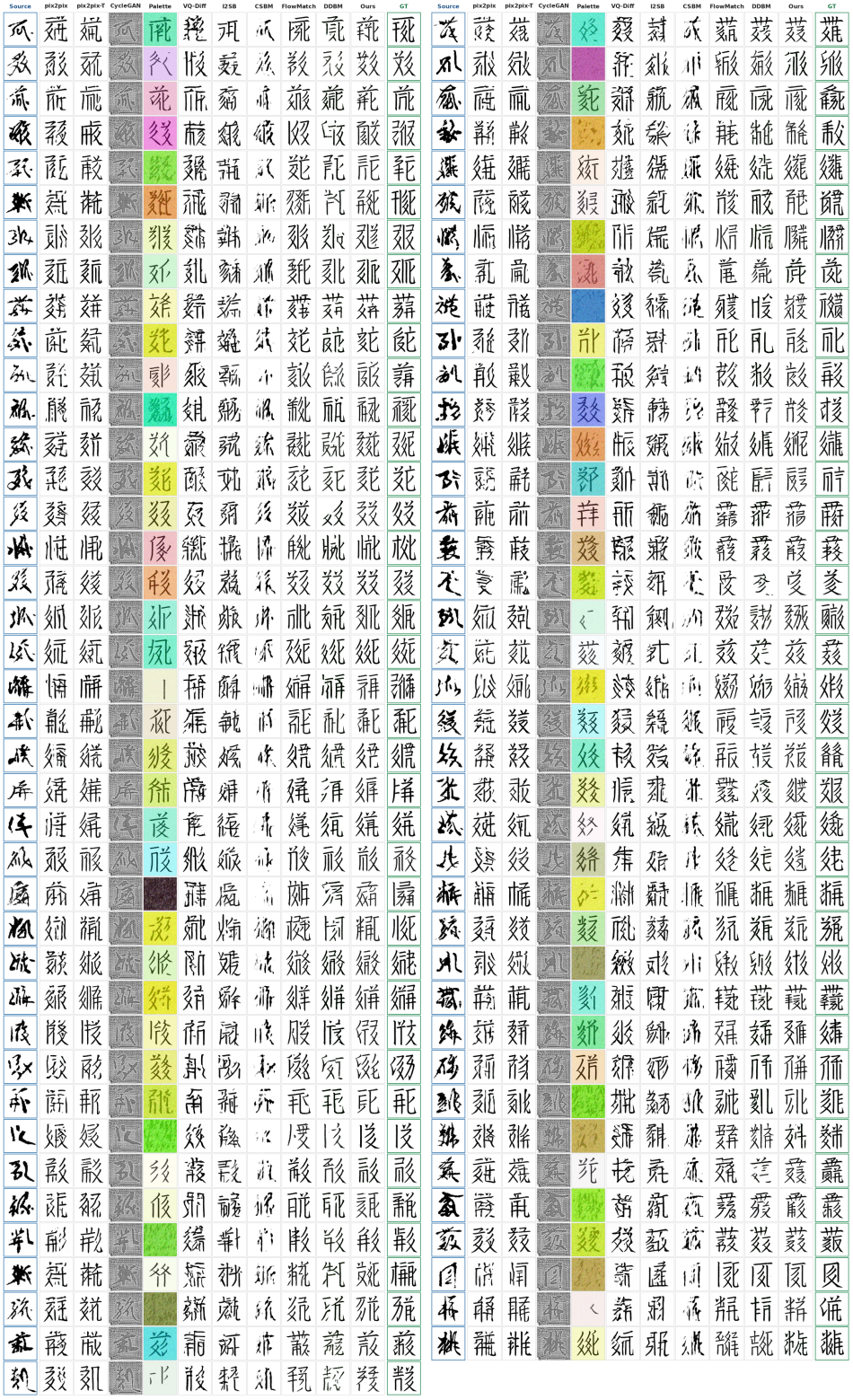

Figure 1: Qualitative comparison under the Unseen Character protocol.

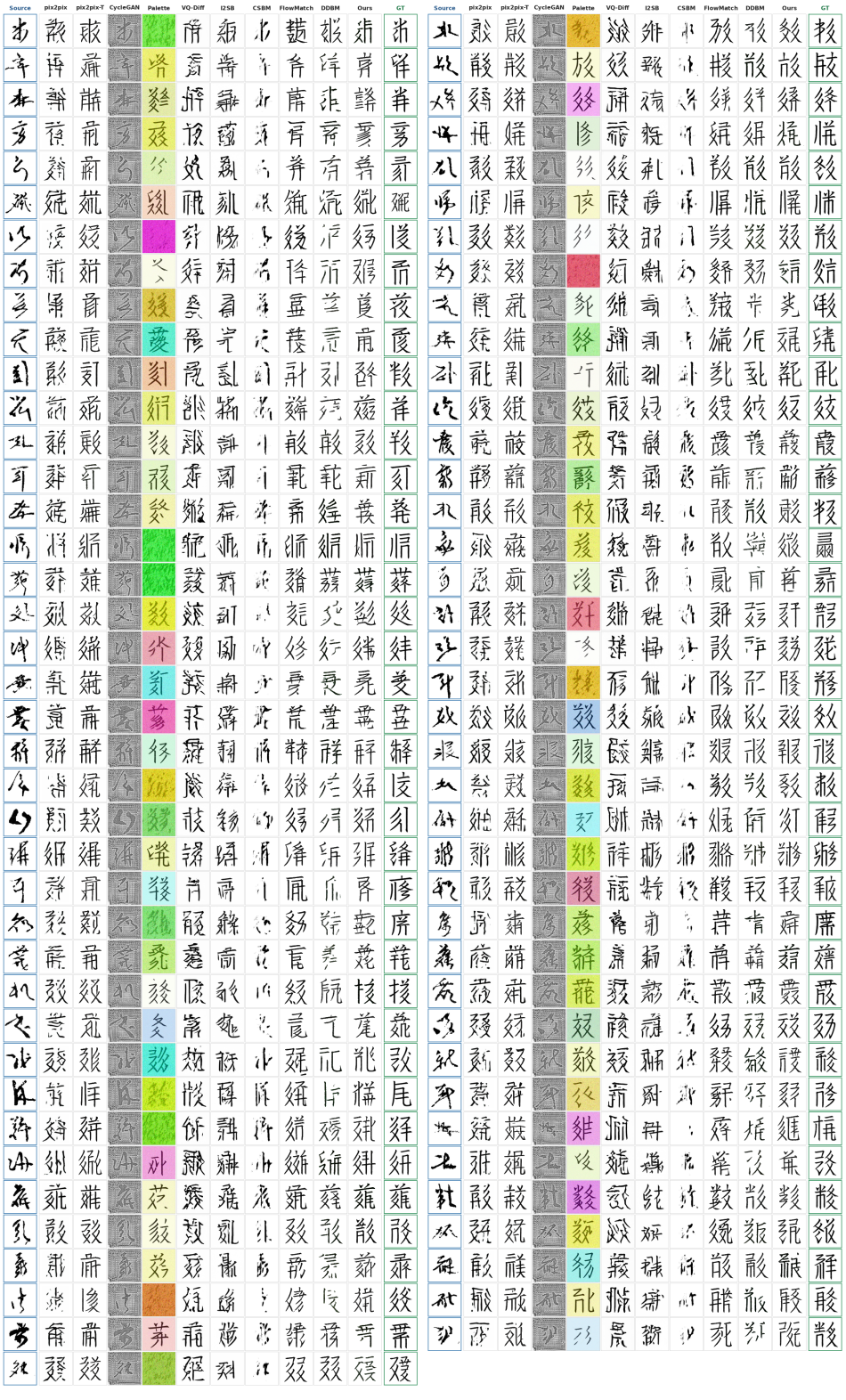

Figure 2: Qualitative comparison under the Seen Character / Unseen Instance protocol.

Table 3: Complete quantitative results. NSCD is normalized by the $128\sqrt{2}$ image diagonal. Mean $\pm$ standard deviation; best is **bold**, second best is underlined.

Method	NSCD ↓	Foreground IoU ↑	SSIM ↑	LPIPS ↓
(a) Unseen Character				
Pix2Pix	0.0182±0.0030	0.2486±0.0615	**0.6203±0.0482**	0.3508±0.0510
CycleGAN	0.03018±0.00206	0.1228±0.0002	0.0079±0.0000	1.0148±0.0009
Pix2Pix-Turbo	0.03677±0.00687	0.2487±0.0032	0.5238±0.0039	0.2969±0.0015
Palette	0.02592±0.01072	0.1504±0.0791	0.2647±0.1589	0.5742±0.2522
VQ-Diffusion	0.02137±0.00164	0.1975±0.0018	0.4870±0.0023	0.3039±0.0019
Flow Matching	0.03388±0.00560	0.2561±0.0036	0.5224±0.0040	**0.1596±0.0011**
DDBM	0.03647±0.01541	0.2864±0.0074	0.5044±0.0074	0.3034±0.0030
I^2SB	0.08465±0.02302	0.2052±0.0016	0.5934±0.0014	0.2234±0.0005
CSBM	0.05264±0.01681	0.1545±0.0023	0.5335±0.0025	0.4300±0.0018
SC-GTM	**0.01602±0.00226**	**0.2919±0.0065**	0.5377±0.0059	0.2475±0.0033
(b) Seen Character / Unseen Instance				
Pix2Pix	0.0212±0.0044	0.2107±0.0713	**0.6573±0.0336**	0.3706±0.0513
CycleGAN	0.03608±0.00320	0.1059±0.0001	0.0107±0.0000	1.0377±0.0008
Pix2Pix-Turbo	0.04053±0.00747	0.2216±0.0029	0.5571±0.0017	0.2975±0.0013
Palette	0.02640±0.00813	0.1550±0.0709	0.3149±0.1762	0.5364±0.2291
VQ-Diffusion	0.02349±0.00283	0.1854±0.0021	0.5323±0.0020	0.3220±0.0020
Flow Matching	0.03959±0.00975	0.2213±0.0042	0.5504±0.0031	**0.1726±0.0012**
DDBM	0.04632±0.04220	**0.2420±0.0150**	0.5465±0.0051	0.3491±0.0069
I^2SB	0.09654±0.03877	0.1847±0.0017	0.6407±0.0007	0.2332±0.0007
CSBM	0.06924±0.09357	0.1246±0.0037	0.6053±0.0012	0.4584±0.0038
SC-GTM	**0.02027±0.00402**	0.2315±0.0058	0.5446±0.0036	0.2809±0.0030

define, for $0 < \kappa < 1$,

$$A^0(i,j) = \begin{cases} \kappa w_{ij}/d_{\max}, & i \neq j, \\ 1 - \sum_{k \neq i} A^0(i,k), & i = j. \end{cases} \tag{15}$$

The kernel is doubly stochastic and has positive self-loops. If $\sum_a \mu_t(i,a) = 1/N$, then $F_t^{\mathrm{ref}}(i,a,j) = \mu_t(i,a)A^0(i,j)$ satisfies both constraints in Eq. (4); hence the transport feasible set is nonempty. For any feasible F_t and row-stochastic rewriting kernel B_t,

$$\sum_b p_{t+1}(j,b) = N \sum_{i,a} F_t(i,a,j) \sum_b B_t(j,a,b) = 1, \tag{16}$$

so validity of the position-wise probability map is preserved.

The generalized KL divergence is

$$D_{\mathrm{gen}}(F\|W) = \sum_{r,j} \left[F_{rj} \log \frac{F_{rj}}{W_{rj}} - F_{rj} + W_{rj} \right], \tag{17}$$

where $r = (i,a)$ indexes a source position–category row. On the active support, the projection has the scaling form

$$F = \mathrm{diag}(u) W \mathrm{diag}(v), \qquad u_r \leftarrow \frac{\mu_r}{(Wv)_r}, \quad v_j \leftarrow \frac{1/N}{(W^\top u)_j}. \tag{18}$$

Rows with zero mass are fixed to zero and excluded from scaling; in implementation the updates are evaluated in the log domain and marginal residuals are checked after the finite number of iterations.

B.2 BOUNDED REWRITING

Let $U(a,b) = 1/K$. To make zero rewriting-control scores correspond to the fixed reference kernel in Eq. (10), define

$$C^0 = \left(1 - \frac{\eta}{\rho}\right) I + \frac{\eta}{\rho} U, \tag{19}$$

so that $(1-\rho)I + \rho C^0 = B^0$. If $s_t^e(j,a,b)$ is the learned local rewriting score, a positive-support parameterization is

$$C_t(j,a,b) = \frac{C^0(a,b)\exp(s_t^e(j,a,b)/\tau_e)}{\sum_{b'} C^0(a,b')\exp(s_t^e(j,a,b')/\tau_e)}. \tag{20}$$

For every row, $B_t(j,a,a) \geq 1-\rho$ and $\sum_{b\neq a} B_t(j,a,b) \leq \rho$. Thus at most a mass ρ leaves its current category in one rewriting step. For any row distribution u,

$$\|uB - u\|_1 = \rho\|uC - u\|_1 \leq 2\rho. \tag{21}$$

For a fixed row-stochastic C and $B = (1-\rho)I + \rho C$, $\|zC\|_1 \leq \|z\|_1$ for any vector z. Setting $z = u - v$ gives

$$\|uB - vB\|_1 \geq (1-\rho)\|u-v\|_1 - \rho\|(u-v)C\|_1 \geq (1-2\rho)\|u-v\|_1. \tag{22}$$

This bound assumes the same fixed rewriting kernel for both inputs; it does not imply global injectivity of the adaptive network when different inputs produce different C_t.

B.3 PATH DISTRIBUTION AND LOSS DECOMPOSITION

For positive-mass rows define the transport kernel

$$A_t(j \mid i,a) = \frac{F_t(i,a,j)}{\mu_t(i,a)}. \tag{23}$$

Split each restoration round into a transport and a rewriting substep,

$$K_{2t}^\theta((i,a),(j,b)) = A_t(j \mid i,a)\mathbf{1}\{b=a\}, \tag{24}$$

$$K_{2t+1}^\theta((i,a),(j,b)) = \mathbf{1}\{j=i\}B_t(i,a,b). \tag{25}$$

With $L = 2T$, forward recursion defines

$$P_\theta^X(s_{0:L}) = \mu_0(s_0)\prod_{\ell=0}^{L-1} K_\ell^\theta(s_\ell, s_{\ell+1}). \tag{26}$$

Replacing controlled kernels by A^0 and B^0 gives R^X with the same initial distribution. Hence

$$\mathrm{KL}(P_\theta^X \| R^X) = \sum_{\ell=0}^{L-1}\sum_{s,s'} m_\ell(s)K_\ell^\theta(s,s')\log\frac{K_\ell^\theta(s,s')}{K_\ell^0(s,s')}, \tag{27}$$

which yields Eq. (12) after substituting the transport edge mass $F_t(i,a,j)$ and rewriting edge mass $\widetilde{\mu}_t(j,a)B_t(j,a,b)$. Zero-mass terms use the convention $0\log 0 = 0$.

The endpoint cross-entropy is also a KL divergence. Let

$$\nu_Y(i,a) = \frac{1}{N}\mathbf{1}\{a = Y_i\}, \qquad \mu_T(i,a) = \frac{p_T(i,a)}{N}. \tag{28}$$

Then

$$\mathrm{KL}(\nu_Y \| \mu_T) = -\frac{1}{N}\sum_i \log p_T(i,Y_i) = \mathcal{L}_{\mathrm{end}}. \tag{29}$$

The reverse KL is not suitable because ν_Y is zero on non-target categories.

B.4 FINITE-STEP REWRITING AND SOFT ENDPOINT SUPERVISION

Define the global category mass $h_t(a) = \sum_i \mu_t(i, a)$. Transport does not change category totals, and the diagonal lower bound of the rewriting kernel gives

$$h_{t+1}(a) \geq (1-\rho)h_t(a), \qquad h_T(a) \geq (1-\rho)^T h_0(a). \tag{30}$$

Therefore an exactly one-hot target endpoint is not guaranteed to be reachable in finitely many bounded steps. Nevertheless, a feasible comparison process that keeps transport at the original positions and redirects all rewritable mass toward Y_j yields

$$p_T^\dagger(j, b) = (1-\rho)^T \mathbf{1}\{b = X_j\} + \left[1 - (1-\rho)^T\right] \mathbf{1}\{b = Y_j\}, \tag{31}$$

so

$$\mathcal{L}_{\text{end}}^\dagger \leq -\log\left(1 - (1-\rho)^T\right). \tag{32}$$

This construction is only an expressivity argument; it is not an intermediate teacher trajectory used in training.

B.5 DYNAMIC CORRESPONDENCE AND REGIONAL PROVENANCE

Let

$$M^X = K_0^\theta K_1^\theta \cdots K_{L-1}^\theta. \tag{33}$$

The model-internal coupling from source position i to terminal position j is

$$\Gamma_{ij}^X = \sum_{a,b} \mu_0(i, a) M^X_{(i,a),(j,b)} = \frac{1}{N} \sum_b M^X_{(i,X_i),(j,b)}. \tag{34}$$

Its row and column marginals are both $1/N$. For a source region I, initialize

$$q_0^I(i, a) = \mu_0(i, a)\mathbf{1}\{i \in I\} \tag{35}$$

and propagate through the already determined forward kernels,

$$q_{\ell+1}^I = q_\ell^I K_\ell^\theta. \tag{36}$$

Summing over categories at the endpoint yields the positional provenance of that region. The controllers should not be rerun for the marked mass, because doing so would define a different trajectory. This coupling is probabilistic provenance inside the model, not a literal stroke genealogy or a causal intervention.