CONSTANT: Towards High-Quality One-Shot Handwriting Generation with Patch Contrastive Enhancement and Style-Aware Quantization Supplementary material

Anh-Duy Le[1], Van-Linh Pham[1], Thanh-Nam Vo[1], Xuan Toan Mai[2], Tuan-Anh Tran[1,2]
[1] Viettel Artificial Intelligence and Data Services Center, Vietnam
[2] Ho Chi Minh City University of Technology, Vietnam
{leanhduy497, phamvanlinh143, thanhnam040501}@gmail.com, {mxtoan, trtanh}@hcmut.edu.vn

A. Introduction

In this supplementary material, we provide more details about Latent Diffusion Models in Sec. B. In Sec. C, we conduct user studies comparing SOTA methods to evaluate visual and style imitation quality. Sec. D provides detailed information about our ViHTGen dataset. We also present the implementation details more clearly in Sec. E. We provide more detail about our evaluation metrics in Sec. F. Next, we show the experimental results on the IIIT-English-Word dataset in Sec. G. Sec. H shows more results between our method and other SOTA methods on multi-language generalization. In Sec. I, we conduct an experiment to evaluate the effectiveness of our method on improving handwritten text recognition performance. We also perform an ablation study on the impact of codebook embedding length on the performance of our model in Sec. J. We provide more detail about the efficiency aspect, including our training cost and inference time, in Sec. K. A discussion about limitations and future work is presented in Sec. L. Finally, we provide more qualitative results of our model on multiple datasets.

B. More Details about LDMs

Diffusion models (DMs) represent a significant advance in generative modeling, often surpassing GANs in various tasks. Starting with DDPM [9], which uses iterative denoising to generate samples, numerous improvements have enhanced quality and control [3, 19, 21]. Techniques like classifier-free guidance (CFG) [8] and multimodal conditioning as seen in GLIDE [13] have further boosted performance, especially in text-to-image generation.

Our diffusion model for handwritten generation is inspired by LDMs [19], which enables the sampling process to occur in the latent space by using a pretrained VAE to compress handwritten image X to a 4-D latent space representation $z \in \mathbb{R}^{4 \times W/8 \times H/8}$. Similar to DDPM [9], the forward process considered a Markov chain of length T consists of gradually adding Gaussian noise to the clean latent representation $\mathbf{z}_0 \sim q(\mathbf{z}_0)$ until it becomes pure noise $p(\mathbf{z}_T) = \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathrm{I})$, with each step defined by a transition probability $q(z_t|z_{t-1})$.

The reverse process aims to learn to undo this noise addition, generating a sample from pure noise by training a Unet model [20] $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{C}, \boldsymbol{X}_s)$ to predict the added noise using a mean-squared error loss:

$$L_{\text{denoising}} = E_{t \sim [1,T], z_0 \sim q(z_0), \epsilon \sim \mathcal{N}(0,\mathbf{I})} \left[\left\| \epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{C}, \boldsymbol{X}_s) \right\|^2 \right] \tag{1}$$

C. User Studies

We perform user studies, including a User Preference Study and User Plausibility Study to better evaluate our methods compared to the others, including One-DM [2], HiGAN+ [6], HWT [1], VATr [17], HiGAN [5], using TypeForm [1] survey platform.

C.1. User Preference Study

We perform sampling a list of 30 text contents using an OOV corpus from IAM [11] dataset and generating 30 images for each method. The test was conducted on 28 participants and received a total of 840 responses. Each participant was asked to choose which image in the given list was the most similar to the real image. As shown in Fig. 1, our method receives the most responses from all participants, with more than 40.6% responses.

C.2. User Plausibility Study

We perform experiment to study whether our method's generated images are indistinguishable from real images. For each question, we ask participants to identify 3 real images in total of 6 images by first showing them 6 examples of real images from the same writers. The study received a total of

[1] https://www.typeform.com/

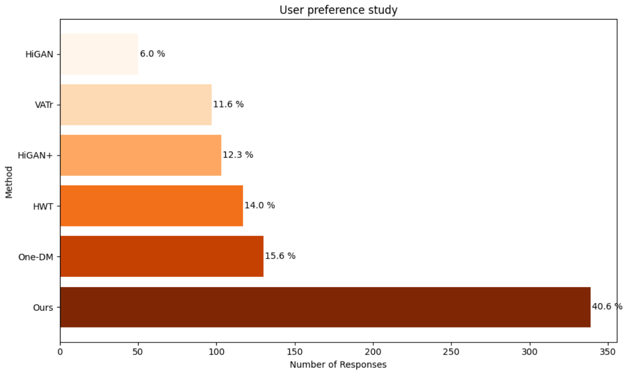


Figure 1. User preference study results

1680 responses from a total of 28 participants. The result is shown in Fig. 2 with the **accuracy of 53.8%**, indicating it is close to random classification.

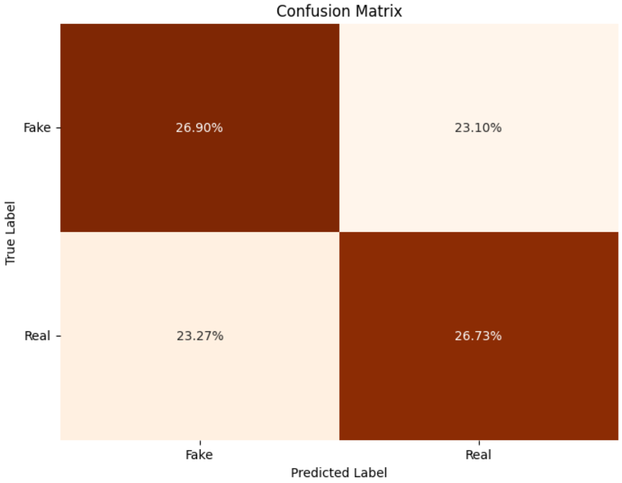


Figure 2. User plausibility study results

D. ViHTGen Dataset

To mimic real-world scenarios, we constructed the **ViHTGen** dataset, featuring diverse handwriting styles on complex backgrounds. The dataset was sourced from over 300 university final exam scripts written by more than 200 individuals. We employed a rigorous semi-automatic annotation pipeline to ensure high-quality labels. First, word instance bounding boxes were automatically extracted using the Google Cloud Vision API[2] and then manually verified. Subsequently, a state-of-the-art Vietnamese vision-language model VinTernVL-1B [4] performed OCR on each word, with all transcriptions being manually checked and corrected to guarantee label accuracy.

After filtering, the final dataset contains over 50,000 word-level images, which we split into a training set of 42,000 and a test set of over 8,000 images. As shown in Fig. 4, we also performed a statistical analysis of text length and character frequency. The resulting dataset is a challenging benchmark for HTG models, featuring a wide variety of stroke styles and complex, noisy backgrounds, as illustrated in Fig. 3. Tab. 1 also shows the differences between our dataset and the standard IAM dataset on many aspects. To encourage reproducibility and foster future research, the ViHTGen dataset will be made publicly available upon publication of this work.

Statistic	ViHTGen (ours)	IAM
Language	Vietnamese	English
# Writers	223	500
# Word Instances	50000+	62857
# Unique words	4644	3332
Style Complexity (Slant, Ink color, Stroke width, character shape)	High	Medium
Background complexity	High	Low
Image source	University Exam	English Sentences

Table 1. Comparison between our ViHTGen and IAM dataset.

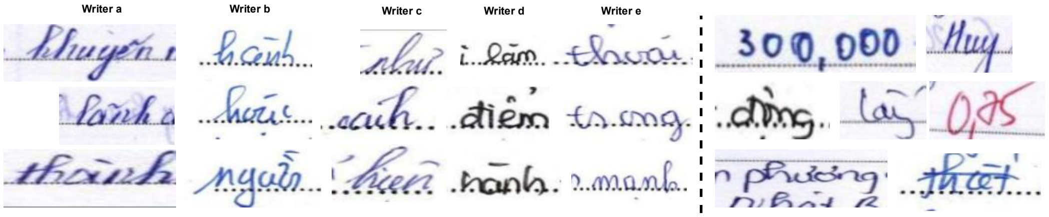
E. More about Implementation Details

E.1. Model Architecture Details

Our model includes three basic components: 1) a Latent Diffusion-based model, 2) a SAQ module for style feature extraction, and 3) a text encoder module.

- **Latent Diffusion-based Models**: We follow the architecture of LDMs [19], which uses a U-Net [20] model that includes a ResNet block followed by a Spatial Transformer block to combine information between context features (here, the style and textual features) and the input features. We follow WordStylist [14] to reduce the number of ResNet blocks to reduce training time, while the number of heads in the Transformer block is set to 4, and the feature dimension is set to 512.
- **SAQ module**: Our proposed SAQ module includes three basic parts: an Inception-V3 backbone, a codebook embedding E, and an AttentionPool fusion module. We chose Inception-V3 for its effectiveness in extracting multiscale features, as is also done in style transfer [7]. The output dimension of Inception-V3 is 768. Since we use a hybrid solution that combines discrete and continuous features as described in the main paper, the output dimension of features $\hat{F}$ in SAQ is 1536. Finally, the At-

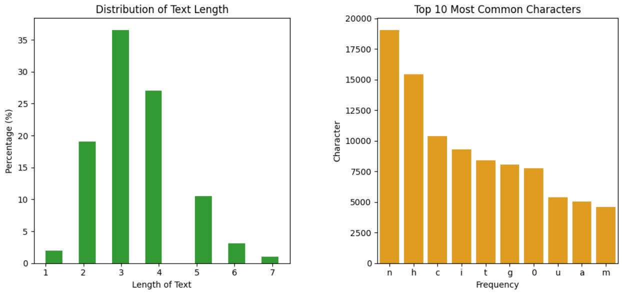
[2] https://cloud.google.com/vision/docs



a) Diversity in style beteen different writers in ViHTGen

b) Complexity in background noise of ViHTGen dataset

Figure 3. An overview of our ViHTGen dataset. a) The diversity of handwritten style between different writers. b) The complexity of our dataset in both background and stroke shape.



(a) The distribution of text length

(b) Most frequent character

Figure 4. Statistical analysis for the ViHTGen dataset. (a) The percentage of images with different text lengths. (b) Frequency distribution of top-10 characters in the dataset

tentionPool module, for which we were inspired by the CLIP framework [18], is used to better fuse information from both discrete and continuous features through a self-attention operation. The final output dimension is projected to 512 through a linear layer before being passed to the LDM.

- **Text encoder module**: The text encoder is a 3-layer Transformer block; each block consists of an MLP block and a self-attention block with a dimension of 512.

E.2. $L_{LatentPCE}$ Implementation Details

We design our $L_{LatentPCE}$ objective as a multi-scale contrastive loss that operates at three distinct scales. At each scale, we extract up to 256 patches of sizes 2×2, 4×4, and 8×8, respectively. These extracted patches are then flattened and projected into a 256-dimensional embedding space using a shallow MLP. The final $L_{LatentPCE}$ is computed as the average of the contrastive losses from each of the three scales.

F. More Details About Evaluation Metrics

F.1. OCR-based Metric

To evaluate the readability of the generated images, we follow the setup from [16]. The core idea is to train an Optical Character Recognition (OCR) model on the generated images and then test its performance on real images. A successful HTG model should produce samples that enable the OCR model to achieve a low Word Error Rate (WER) on real data. We use the sequence-to-sequence OCR architecture from [10]. For the evaluation, we first train the OCR model on images generated for the IAM test set. The model is trained for 200,000 iterations with a batch size of 128. After training, we test the OCR model on the real IAM test set and calculate the WER.

F.2. Writer Classification Metric

Besides evaluating readability, we also assess the style imitation ability using a writer classifier model. This evaluation strategy has been used in previous works [14, 15]. In our work, we use a ResNet18 model pre-trained on ImageNet as the base architecture. To train the classifier, we split the real IAM test set into an 80/20 ratio for training and validation. The trained model is then used to evaluate the writer classification accuracy on the generated version of IAM test set, which serves as a measure of our HTG model's style imitation capability.

G. Experiments on IIIT-English-Word Dataset

We perform experiments on IIIT-English-Word dataset [12] to compare the performance between our CONSTANT model and One-DM [2] in terms of FID and HWD score. The quantitative result show in Table 2 and the qualitative result show in Fig. 5

Method	HWD ↓	FID ↓
One-DM [2]	1.22	17.74
Ours	**0.73**	**10.22**

Table 2. Quantitative results on IIIT-English-Word test set.

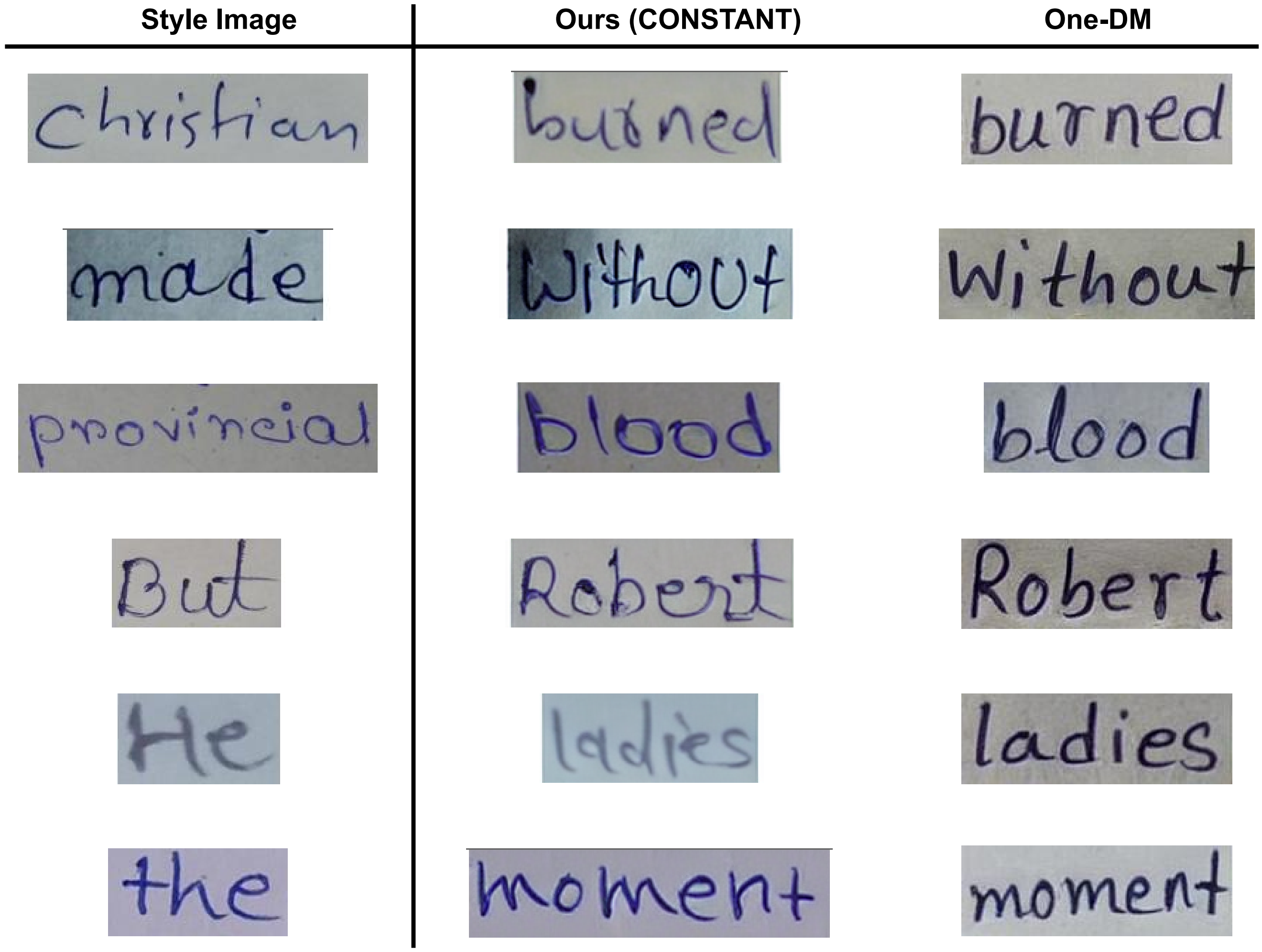


Figure 5. Comparisons between our method and One-DM [2] on IIIT-English-Word.

H. More Results on Multi-language Generalization

To better demonstrate the effectiveness of our method on other languages, we perform additional experiments comparing our method to a state-of-the-art (SOTA) competitor. In this experiment, we use DiffusionPen [15] as the baseline model. Since DiffusionPen is a few-shot model, we train it in a one-shot setting to ensure a fair comparison. Similar to the evaluation against One-DM in the main paper, we compare our method with DiffusionPen on both Chinese and Vietnamese datasets. As shown in Tab. 3 and Fig. 6, our method achieves better results on both languages in terms of visual quality and style adaptation ability compared to DiffusionPen. This also shows the robustness of our method in the one-shot setting, whereas the DiffusionPen model's performance is low when adapted to this setting (e.g., achieving a Chinese FID score of 54.07 compared to our 22.74).

	Chinese		Vietnamese	
Method	HWD ↓	FID ↓	HWD ↓	FID ↓
DiffusionPen [15]	0.57	54.07	1.05	23.59
Our	**0.37**	**22.74**	**0.83**	**18.81**

Table 3. Quantitative comparisons with One-DM on Chinese and Vietnamese scripts in terms of FID and HWD.

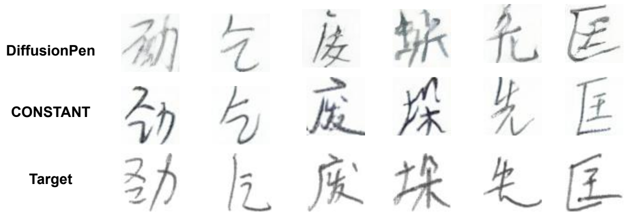


Figure 6. Qualitative results between CONSTANT and DiffusionPen [15] in Chinese script.

I. Handwritten Text Recognition Improvement

We conduct experiments to evaluate the improvement of HTR model when increasing the number of generated data in the training set. The experiments include training on the real IAM training set and gradually increasing the number of handwritten images generated using our methods by 100K each time.

Data Source	Accuracy
Real data	81.76
Real data + 100K	82.13
Real data + 200K	**82.96**

Table 4. Generated data helps improve OCR performance on IAM test set

J. Ablation on Codebook Embedding Length

We conducted an ablation study to investigate the impact of the codebook embedding length in SAQ and to validate our hypothesis regarding its correlation with dataset complexity. The study involved varying the codebook size (K) across three datasets with distinct style complexities: the relatively simple IAM dataset and the more visually complex IMGUR5K and IIIT-English-Word datasets. We evaluated codebook sizes of K=512, K=1024, and K=2048. As shown in Table 5, the results align with our hypothesis. For the simple IAM dataset, the largest codebook (K=2048) slightly worsened performance (FID 14.10, HWD 0.86), which suggests overfitting. In contrast, for the more intricate styles of IMGUR5K and IIIT-English-Word, the larger K=2048 codebook yielded the best performance, enabling the model to capture a richer diversity of style features. These results effectively consolidate the hypothesis stated in our main paper.

	IAM		IMGUR5K		IIIT-English-Word	
Codebook length	HWD ↓	FID ↓	HWD ↓	FID ↓	HWD ↓	FID ↓
512	0.86	13.80	1.02	13.37	0.75	14.02
1024	**0.84**	**12.46**	1.03	12.9	0.77	12.61
2048	0.86	14.10	**0.99**	**11.48**	**0.73**	**10.22**

Table 5. Ablation Study About The Codebook Embedding Length in SAQ Module

K. Effiency Analysis

To provide further details on the training cost and efficiency of our method, we compare it with One-DM in terms of training cost. Table 6 provides detailed information on the training and inference costs between our method and the baseline One-DM model. Experiments for both methods were conducted on the same NVIDIA V100 machine.

Criteria	CONSTANT	One-DM
Model parameters	124×10^6	185×10^6
Training VRAM	6.82GB	18.66GB
Inference time[3]	1.25 s/sample	1.85 s/sample
Inference VRAM	2.1GB	2.9GB

Table 6. Comparison training and inference cost between CONSTANT and One-DM.

L. Limitation and Future Works

Despite achieving state-of-the-art results, our method has limitations. The model's style extraction can be compromised when reference images are excessively blurry or feature highly complex backgrounds (Fig. 7a). Similarly, for overly intricate or nearly illegible handwriting, our model may prioritize content readability over precise style imitation, as shown in Fig. 7. Another limitation is that the codebook size is determined empirically, which may not be optimal for datasets with different style complexities. Future work will focus on improving style extraction for highly artistic text and exploring methods for generating longer lines of text.

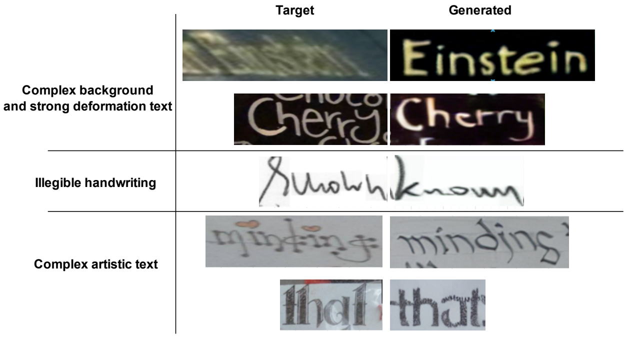


Figure 7. Visualization of some failure cases.

M. More Qualitative Results

References

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Mubarak Shah. Handwriting transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1086–1094, 2021. 1

[2] Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-dm: One-shot diffusion mimicker for handwritten text generation. In *European Conference on Computer Vision*, pages 410–427. Springer, 2025. 1, 3, 4

[3] Sampling taken under 50 steps

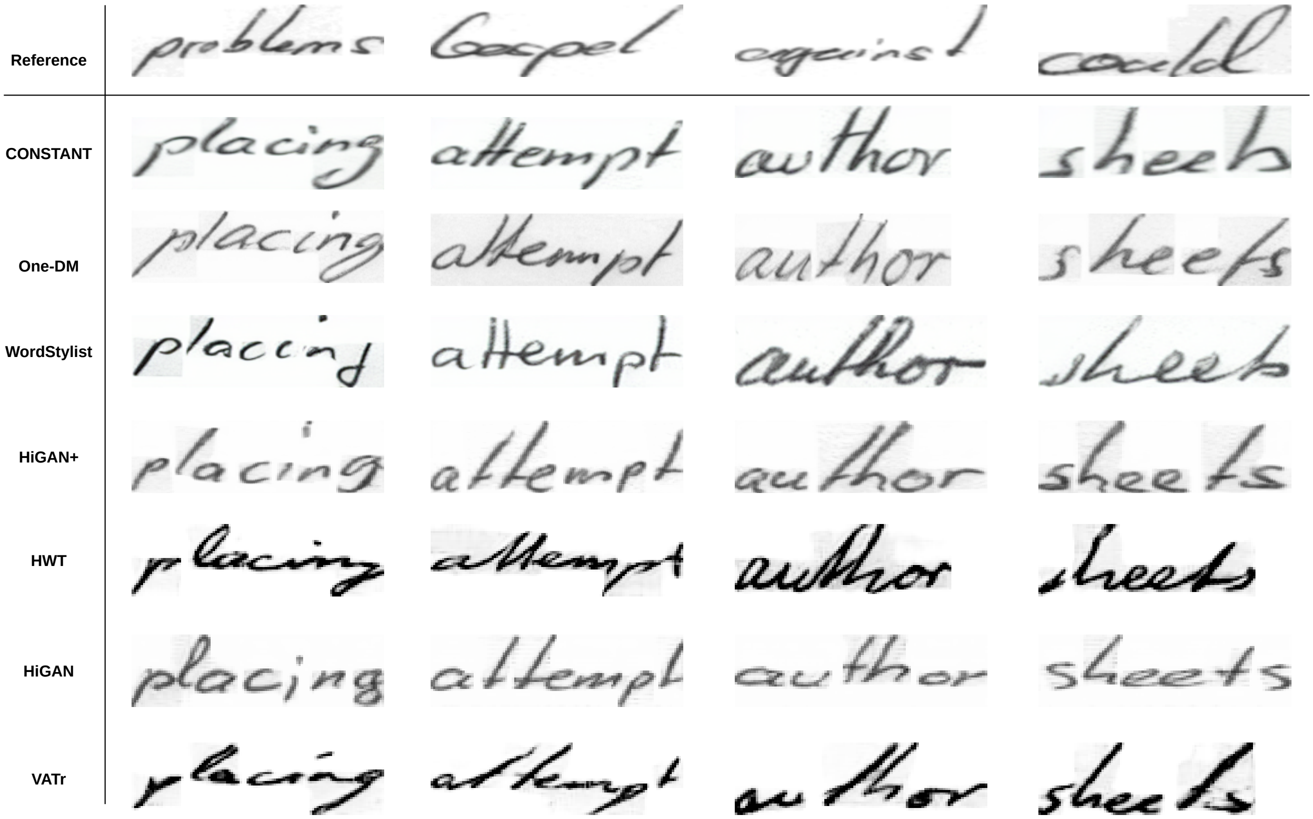


Figure 8. Visualization of arbitrary textual content between different methods on IAM test dataset.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[4] Khang T Doan, Bao G Huynh, Dung T Hoang, Thuc D Pham, Nhat H Pham, Quan Nguyen, Bang Q Vo, and Suong N Hoang. Vintern-1b: an efficient multimodal large language model for vietnamese. *arXiv preprint arXiv:2408.12480*, 2024. 2

[5] Ji Gan and Weiqiang Wang. Higan: Handwriting imitation conditioned on arbitrary-length texts and disentangled styles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7484–7492, 2021. 1

[6] Ji Gan, Weiqiang Wang, Jiaxu Leng, and Xinbo Gao. Higan+: Handwriting imitation gan with disentangled representations. *ACM Trans. Graph.*, 42(1), 2022. 1

[7] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*, 2017. 2

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 1

[10] Dmitrijs Kass and Ekta Vats. Attentionhtr: Handwritten text recognition based on attention encoder-decoder networks. In *International Workshop on Document Analysis Systems*, pages 507–522. Springer, 2022. 3

[11] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5:39–46, 2002. 1

[12] Ajoy Mondal, Krishna Tulsyan, and CV Jawahar. Bridging the gap in resource for offline english handwritten text recognition. In *International Conference on Document Analysis and Recognition*, pages 413–428. Springer, 2024. 3

[13] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1

[14] Konstantina Nikolaidou, George Retsinas, Vincent Christlein, Mathias Seuret, Giorgos Sfikas, Elisa Barney Smith, Hamam Mokayed, and Marcus Liwicki. Wordstylist: styled verbatim handwritten text generation with latent diffusion models. In *International Conference on Document Analysis and Recognition*, pages 384–401. Springer, 2023. 2, 3

[15] Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. Diffusionpen: Towards controlling

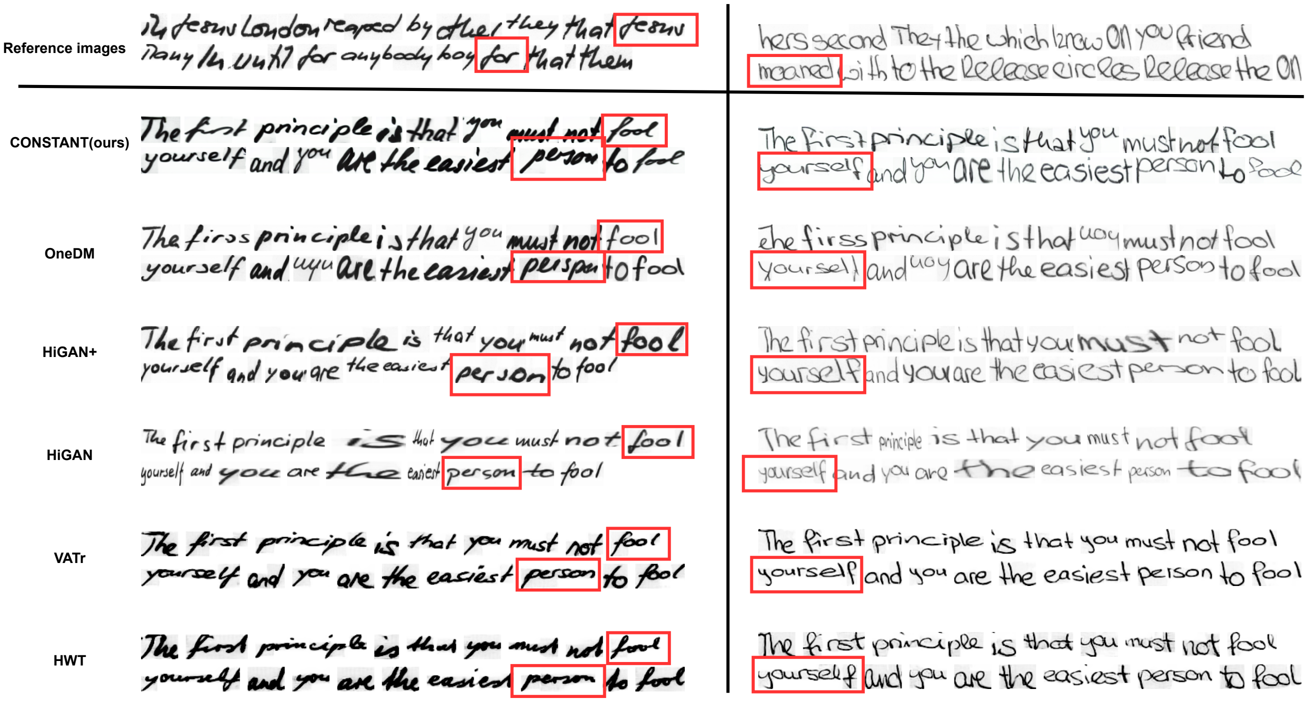


Figure 9. Visualization of our CONSTANT with other competitors, where RED box shows the style adaptability between different methods to the reference image.

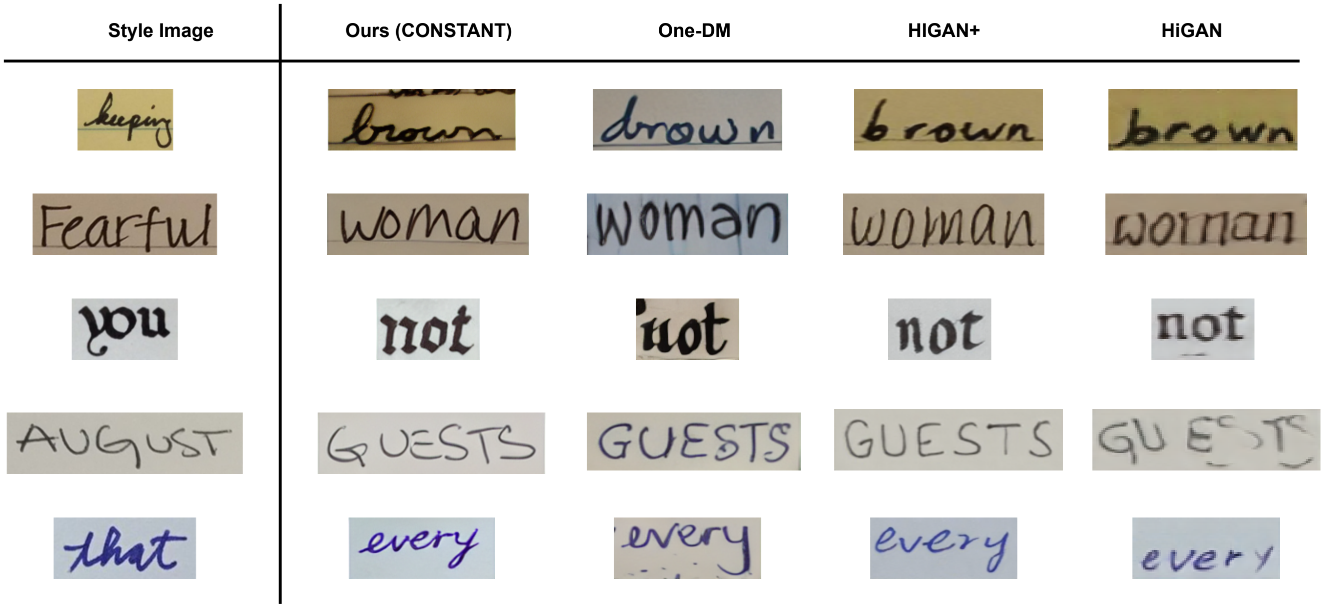


Figure 10. Other examples on IMGUR5K dataset.

the style of handwritten text generation. *arXiv preprint arXiv:2409.06065*, 2024. 3, 4

[16] Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. Rethinking htg evaluation: Bridging generation and recognition. *arXiv preprint arXiv:2409.02683*, 2024. 3

[17] Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Handwritten text generation from visual archetypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22458–22467, 2023. 1

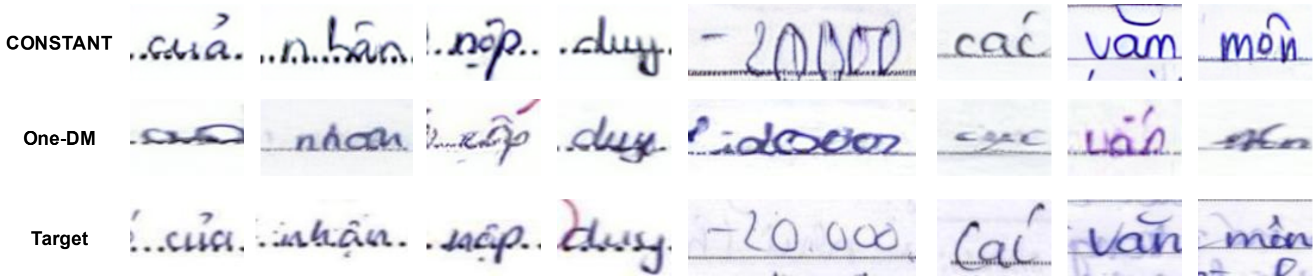


Figure 11. Other examples on ViHTGen dataset.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1, 2

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1