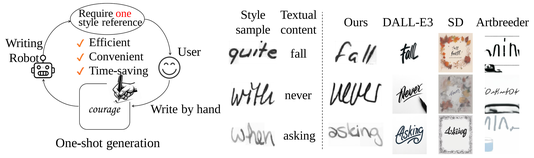
One-DM: One-Shot Diffusion Mimicker for Handwritten Text Generation

Gang Dai*, Yifan Zhang*, Quhui Ke, Qiangya Guo, Shuangping Huang
Email: eedaigang@mail.scut.edu.cn, yifan.zhang@u.nus.edu Code: https://github.com/dailenson/One-DM

One-shot Generation

One-shot handwritten text generation is to imitate user's writing style from only **a single reference**, and generate stylized handwritten text images of any content



Limitations of Existing Methods

- Most require users to provide a few reference samples (*typically 15*), making them inconvenient to use
- Some achieve one-shot generation but perform poorly in emulating handwriting styles due to their simple style encoder design (e.g., vanilla *CNN or transformer*)

Motivations

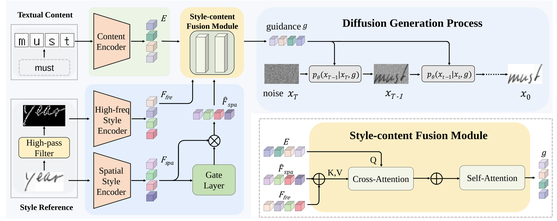
- Accurately extracting writing style from only one reference sample is non-trivial
- Background noise is commonly present in style samples, which further increases the difficulty of style extraction

Drawing from the high-frequency information in samples that clearly showcase style patterns, we incorporate it to enhance style extraction. Additionally, we adaptively remove harmful noise to further improve style learning

Method Overview

One-DM consists of a **style-enhanced module**, a **gate mechanism**, and a **conditional diffusion model**

- **The style-enhanced module** seeks to independently extract **spatial and high-frequency style features** from reference sample and its high-frequency information
- The gate mechanism **selectively filters out** background noise from the reference style features, allowing only **meaningful style patterns** to pass



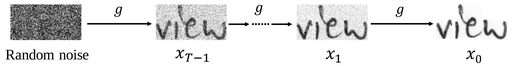
Conditional Diffusion Model

- Use content feature E as the query to adaptively aggregate the information in two learned style features F_{spa} and F_{fre}. The aggregated results and E are then fed into a self-attention to obtain a merged condition g

$$O = Atten_1(Q_1 = E, K_1 = V_1 = F_{spa} + F_{fre}),$$

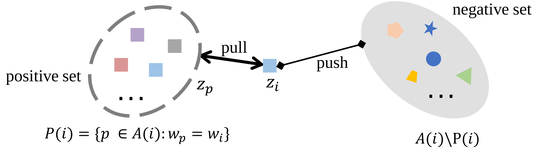
$$g = Atten_2(Q_2 = K_2 = V_2 = O + E).$$

- The diffusion model progressively synthesizes the desired handwritten text images x_0 conditioned on g, starting from a random noise



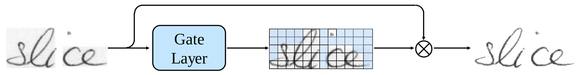
One-shot Style Learning

❶ **Laplacian style extraction**: Use **Laplacian kernel** to extract high-frequency components H_s from one-shot sample, then **align** H_s **from the same writer**



$$\min -\frac{1}{N}\sum_{i\in M}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\mathrm{sim}(z_i,z_p)/\tau)}{\sum_{a\in A(i)}\exp(\mathrm{sim}(z_i,z_a)/\tau)}$$

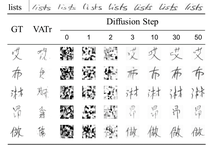
❷ **Background noise suppression**: Extracted spatial style features are fed into **gate layer** to obtain gate units, allowing for a **higher pass rate** for informative style features



Analysis

- Style-enhanced module improves the style extraction, while gate mechanism reduces the background noise
- One-DM can generate **new styles** based on seen styles
- One-DM first construct content then refine style

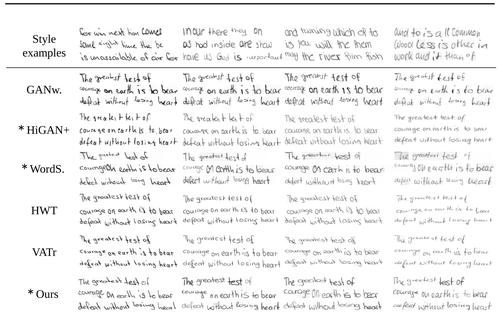
Base $\mathcal{E}_{fre}$	Gate	Style samples	FID ↓
✓		cloudfoul	108.44
✓	✓	cloudfoul	105.73
✓	✓	cloudfoul	104.52
✓	✓	✓ cloudfoul	102.75



Handwritten Text Generation

- One-DM **outperforms** previous handwritten text generation methods that require **15x more references**

Method	Shot	Styled Evaluation				Style-agnostic	
		IV-S	IV-U	OOV-S	OOV-U	FID↓	GS↓
TS-GAN [9]	One	118.56	128.75	127.11	136.67	20.65	4.88×10^{-2}
GANwriting [27]	Few	120.07	124.30	125.87	130.68	28.37	5.67×10^{-2}
HiGAN+ [15]	One	117.33	116.95	121.55	121.48	22.95	2.06×10^{-2}
GC-DDPM [12]	One	99.86	105.73	112.52	118.39	19.05	1.31×10^{-2}
WordStylist [44]	One	98.10	104.27	109.45	115.52	18.58	2.85×10^{-2}
HWT [5]	Few	109.25	106.90	116.55	113.52	18.99	4.41×10^{-3}
VATr [46]	Few	103.75	101.73	111.64	108.76	16.03	1.74×10^{-2}
Ours (One-DM)	One	**89.47**	**98.36**	**93.30**	**102.75**	**15.73**	$\mathbf{1.98\times10^{-3}}$



Application to Other Languages

- One-DM can generate handwritten scripts in **different languages** well