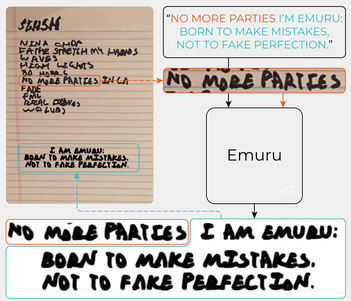
Zero-Shot Styled Text Image Generation, but Make It Autoregressive

Vittorio Pippi*, Fabio Quattrini*, Silvia Cascianelli, Alessio Tonioni, Rita Cucchiara

Google · AImage Lab UNIMORE Università degli Studi di Modena e Reggio Emilia · CVPR Nashville June 11-15, 2025

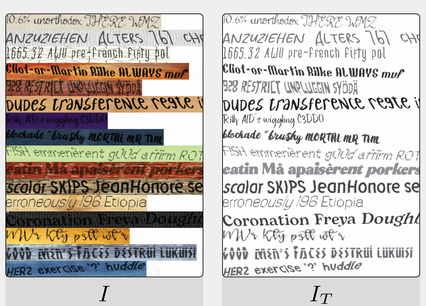
Overview

Emuru is the first autoregressive text-image generator that mimics any handwriting / font in zero-shot and outputs lines of arbitrary length without background artifacts.



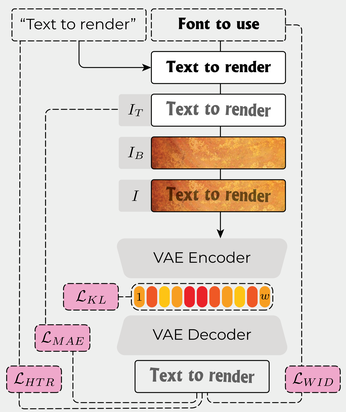
Synthetic Dataset

Emuru has never seen a single real image of text in training. Instead, it was trained on >2.2M lines rendered in >100k calligraphy and typography on random paper textures



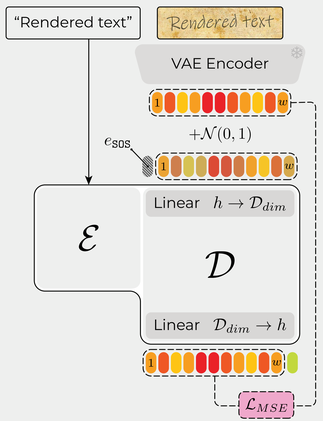
Train the VAE

A lightweight β-VAE turns each 64 px-high line into an 8 × w latent grid and is trained with L1 + β KL, writer-ID, and HTR auxiliaries to isolate stroke style and discard background.



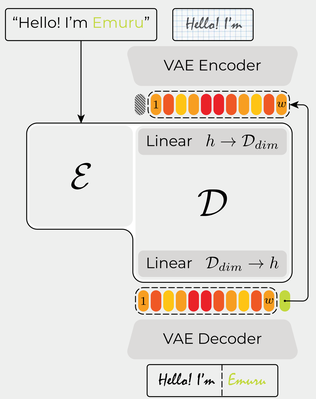
Train Emuru

A T5-Large encoder–decoder learns to autoregress latents: noisy teacher-forcing with MSE, then curriculum from 4-7-word to 32-word extends line lenght limits.



Inference time!

Feed one reference line + desired text; the Transformer emits latents until "padding" latents appear (detected via the t-SNE cluster), then the VAE decodes the final styled image.



Visual prompting

Emuru can fully exploit the style image thanks to the paired visual-textual input it receives. Thus, it can imitate unusual shapes for characters if they appear in the input.

Results

IAM Words	FID↓	ΔCER↓	HWD↓
VATr++	31.91	0.07	2.54
DiffPen	**15.54**	0.06	**1.78**
Emuru	63.61	0.19	3.03

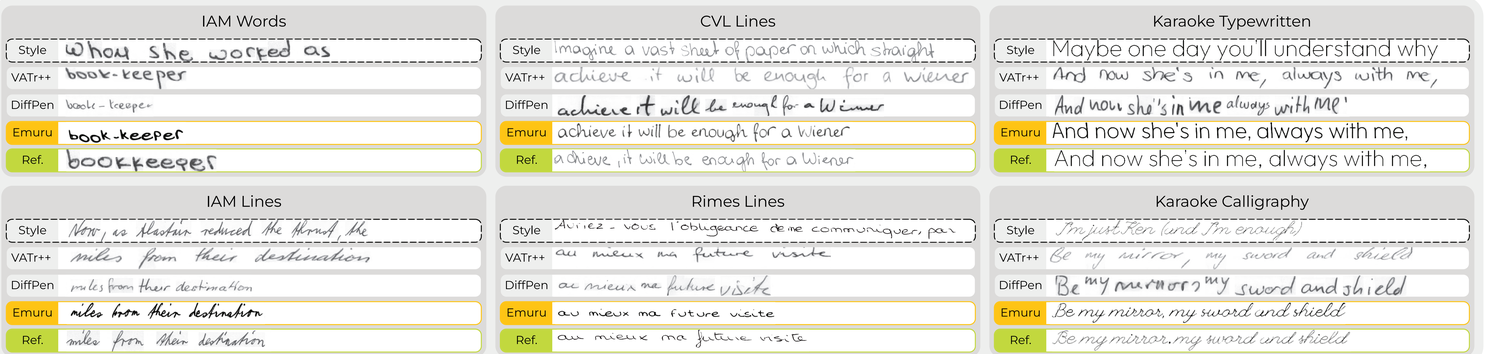
IAM Lines	FID↓	ΔCER↓	HWD↓
VATr++	34.00	0.03	2.38
DiffPen	**12.89**	0.03	2.13
Emuru	13.89	0.14	**1.87**

CVL Lines	FID↓	ΔCER↓	HWD↓
VATr++	35.53	0.12	2.18
DiffPen	40.40	**0.01**	2.99
Emuru	**14.39**	0.13	**1.82**

RIMES Lines	FID↓	ΔCER↓	HWD↓
VATr++	110.04	0.10	2.83
DiffPen	89.79	**0.04**	2.58
Emuru	**26.93**	0.25	**2.18**

Karaoke Calligraphy	FID↓	ΔCER↓	HWD↓
VATr++	67.16	**0.01**	3.96
DiffPen	34.19	0.16	4.18
Emuru	**13.87**	0.13	**2.24**

Karaoke Typewritten	FID↓	ΔCER↓	HWD↓
VATr++	76.03	**0.01**	4.15
DiffPen	78.07	0.14	4.71
Emuru	**9.85**	0.11	**1.28**



References

1. Nikolaidou et al. *DiffusionPen: Towards Controlling the Style of Handwritten Text Generation.* ECCV, 2024.
2. Vanherle et al. *VATr++: Choose Your Words Wisely for Handwritten Text Generation.* IEEE Trans. PAMI, 2024.

Karaoke · HWD · Font² · Emuru