# Supplementary Material for HWD: A Novel Evaluation Score for Styled Handwritten Text Generation

Vittorio Pippi
vittorio.pippi@unimore.it
Fabio Quattrini
fabio.quattrini@unimore.it
Silvia Cascianelli
silvia.cascianelli@unimore.it
Rita Cucchiara
rita.cucchiara@unimore.it

University of Modena and Reggio Emilia
Modena, IT

## 1 Additional Results on the Sensitivity to Handwriting

In this section, we report the results on the sensitivity of the HWD and the FID to the handwriting style obtained on the Norhand [5] and BanglaWriting [6] multi-author datasets. We consider half of the samples for each featured writer as references and the other half as if they were the output of a perfect Styled HTG model. Then, we compare the distributions of the HWD and the FID values computed on text images of multiple matching and non-matching authors pairs. The obtained distributions are reported in Figure 1.

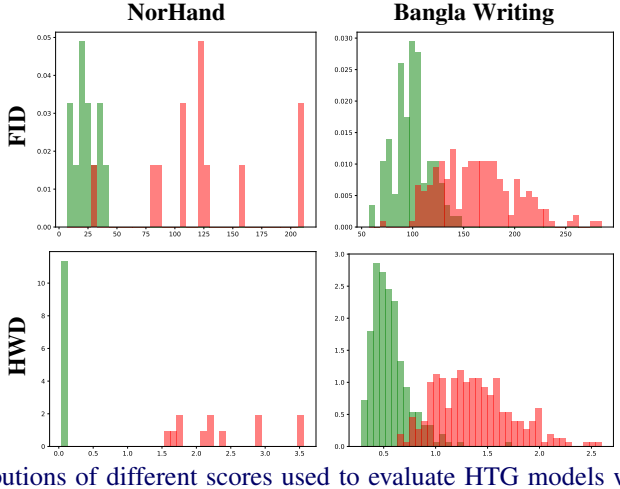



Figure 1: Distributions of different scores used to evaluate HTG models when applied on same-author (green) or different-author (red) subsets. The overlap area is in dark red.

## 2 Further Comparison Between HTG Approaches

In Table 1, we report some qualitative examples of images generated by two HTG approaches being scored with both FID and HWD. These show that HWD better separates cases in which HTG models perform one better than the other, compared to the FID, which has similar values both for good cases and failure cases.

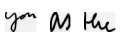
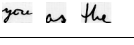
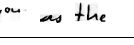
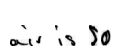
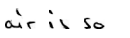
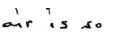
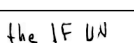
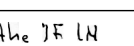
| Reference | VATr | | HWT | |
|---|---|---|---|---|
| | HWD | FID | HWD | FID |
|  |  | |  | |
| | 0.84 | 128.4 | 1.09 | 128.7 |
|  |  | |  | |
| | 0.70 | 111.7 | 0.93 | 111.9 |
|  |  | |  | |
| | 0.60 | 86.6 | 0.92 | 86.7 |

Table 1: Qualitatives, FID, and HWD of HTG models.

## 3 Alternative Metric Distances for HWD

In Table 2 we report the results obtained on the IAM dataset when using Mahalanobis and the Hamming distances in the final step of the HDW computation. It emerges that the Euclidean distance works best for HWD, leading to the smaller Overlap and EER.

| Distance | Overlap | EER |
|---|---|---|
| Mahalanobis | 7.4 | 3.6 |
| Hamming | 4.3 | 2.1 |
| Euclidean | **0.7** | **0.3** |

Table 2: Ablation results when changing the distance metric at the final step of HWD.

## 4 Additional Results on the Sensitivity to the Number of Samples

In this section, we report further results on the numerical stability of the proposed **HWD**, compared to the **FID** and two baseline scores, namely the **FID w/ Euclidean** (obtained by computing the Euclidean distance on the Inception-v3 features) and **HWD w/ Fréchet** (obtained by computing the Fréchet distance on the VGG16). In particular, we use the images from the considered single-author datasets (ICFHR14 [7], Saint Gall [3], Leopardi [1], Rodrigo [8], Washington [4], and LAM [2]). The results are expressed as mean and range between the 25th and 75th percentiles of the values obtained over multiple runs by varying

the number of samples. These are reported in Figure 2. For the plots in each row, we consider the whole indicated dataset as the set of reference images and compute the score when comparing it with a variable number of samples from the other datasets and from the dataset itself for reference.

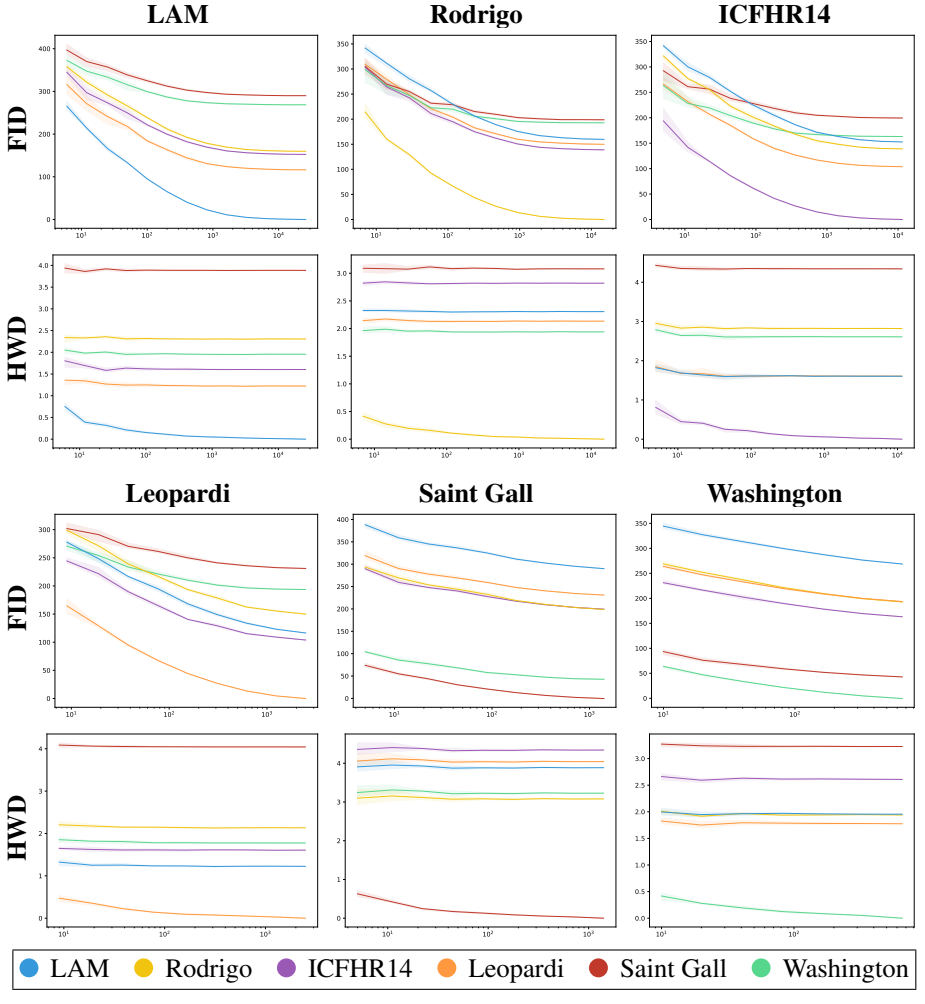



Figure 2: Comparison between FID and HWD with varying number of samples on different single-author datasets. The lines denote the mean, and the transparent bands represent the range between the 25th and 75th percentiles, obtained with 10 calculation runs.

# 5 Computation Time Comparison

We consider the computation time, consisting of image representation and distance computation, of FID and HDW on the same hardware and data. Computing the FID score on 25823/25823 real/fake images from the LAM dataset takes 426.12s + 9.03s (image representation + distance computation), while the computation time of HWD is 135.50s + 0.01s.

# References

[1] Silvia Cascianelli, Marcella Cornia, Lorenzo Baraldi, Maria Ludovica Piazzi, Rosiana Schiuma, and Rita Cucchiara. Learning to Read L'Infinito: Handwritten Text Recognition with Synthetic Training Data. In *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, 2021.

[2] Silvia Cascianelli, Vittorio Pippi, Maarand Martin, Marcella Cornia, Lorenzo Baraldi, Kermorvant Christopher, and Rita Cucchiara. The LAM Dataset: A Novel Benchmark for Line-Level Handwritten Text Recognition. In *ICPR*, 2022.

[3] Andreas Fischer, Volkmar Frinken, Alicia Fornés, and Horst Bunke. Transcription alignment of Latin manuscripts using hidden Markov models. In *HIP*, 2011.

[4] Andreas Fischer, Andreas Keller, Volkmar Frinken, and Horst Bunke. Lexicon-free handwritten word spotting using character HMMs. *Pattern Recognition Letters*, 33(7): 934–942, 2012.

[5] Martin Maarand, Yngvil Beyer, Andre Kåsen, Knut T Fosseide, and Christopher Kermorvant. A comprehensive comparison of open-source libraries for handwritten text recognition in norwegian. In *DAS*, 2022.

[6] Muhammad F Mridha, Abu Quwsar Ohi, M Ameer Ali, Mazedul Islam Emon, and Muhammad Mohsin Kabir. BanglaWriting: A multi-purpose offline Bangla handwriting dataset. *Data in Brief*, 34:106633, 2021.

[7] Joan Andreu Sánchez, Verónica Romero, Alejandro H Toselli, and Enrique Vidal. ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS). In *ICFHR*, 2014.

[8] Nicolás Serrano, Francisco Castro, and Alfons Juan. The RODRIGO Database. In *LREC*, 2010.