DOI: https://doi.org/10.15276/ict.03.2026.09
UDC 004.855.91:514.112

From isolated glyphs to realistic handwriting: diffusion-refined synthetic data generation for Ukrainian cyrillic text recognition

Pavlo Berezin
ORCID: https://orcid.org/0000-0003-1869-5050; pavlo.berezin@gmail.com
National University of Kyiv Mohyla Academy, 2, Skovorody vul. Kyiv, 04070, Ukraine

ABSTRACT

Recognition of handwritten Cyrillic, and Ukrainian in particular, remains constrained by the lack of richly annotated handwriting corpora. This study treats the shortage of training data as the central bottleneck and investigates synthetic data generation as a response. A comparative evaluation of established optical and handwritten text-recognition systems on several Cyrillic handwriting benchmarks confirms that this task remains difficult and motivates a data-centric methodology. The main contribution is an end-to-end pipeline for generating synthetic Ukrainian handwriting. It constructs an isolated glyph bank from a purpose-designed full-alphabet collection form, filters anomalous samples in a learned representation space, extracts cleaned phrase content from a modern Ukrainian corpus, composes multi-word line images, applies conventional visual augmentation, and refines the resulting composites with a prompt-controlled general-purpose image-editing model to obtain more natural handwriting. The work further proposes an automatic verification stage to reduce the risk that visually plausible but textually incorrect samples enter a future training corpus. The architecture can combine glyphs from different writers within a line while preserving the target transcription and allowing controlled refinement of writing appearance. In the current study, the diffusion-refined extension is demonstrated qualitatively rather than through a full downstream training evaluation. The paper therefore presents a controllable architecture for Ukrainian handwriting synthesis, clarifies its limitations, and specifies the validation protocol required to establish its recognition value.

Keywords: Synthetic data generation; handwritten text generation; diffusion models; image-to-image editing; Cyrillic OCR; Ukrainian handwriting;, handwritten text recognition; data augmentation

For citation: **Berezin P. "From isolated glyphs to realistic handwriting: diffusion-refined synthetic data generation for Ukrainian cyrillic text recognition". *Informatics. Culture. Technology*. 2026; Vol. 3 No. 1 (3): 107–119. DOI: https://doi.org/10.15276/ict.03.2026.09**

Introduction. Printed and handwritten documents remain key media for storing and transmitting information. In practice, much of this material exists as unstructured data: scanned pages, photographs, PDFs, and word-processing files. Automatic text processing therefore plays an important role across many domains. In historical research, it enables new forms of retrieval and analysis that support deeper interpretation of archival sources. In contemporary socioeconomic settings, processing everyday documents such as receipts can substantially improve the speed and accuracy of routine tasks. It also improves accessibility by converting text into alternative formats such as audio.

Text document processing is especially important in the context of the Russo-Ukrainian War, where large volumes of documentary evidence must be analyzed quickly and accurately. One practical example is extracting information about a missile: where it was manufactured, who participated in its production, what designation it carries, and who deployed it. Yet Ukrainian and other Cyrillic languages remain less well served by modern text-processing tools than English. Some widely used OCR systems still provide limited support for Ukrainian printed text and even weaker support for handwritten Cyrillic [18].

The performance degradation of Cyrillic OCR models is a derivative of the lack of handwritten datasets. At the same time, synthetic data generation demonstrated outstanding results in tasks of scene text detection and recognition [22], which suggests that realistic handwritten-text synthesis could materially improve OCR performance for Cyrillic.

The lack of datasets, especially the hand-written ones, is the biggest obstacle to the creation of Cyrillic text recognition models. However, this problem could be partially solved by synthetic generation of printed and hand-written data [6]. The rapid development of large language models (LLM) will allow for the expansion of the capabilities of getting such data, although it would still require some adaptation for Cyrillic texts. It is important to note, that one of the biggest problems we would need to face is hallucination of LLMs and checking its results [12].

A text document processing system usually consists of multiple components, which include: text detection on the image, text recognition, pre-processing (e.g. scaling, skew correction, noise removal, rectification, etc.), text analysis, and information retrieval. Another important component is the information extraction test module, which is even more important in cases where mistakes or hallucinations are possible.

Some of these components are language agnostic and can be adapted to any language without much additional changes. Those include text detection modules and data testing. Other components, like text recognition and information extraction, are heavily dependent on language specifics and require additional adaptation to each specific language.

For example, both text detection and data testing can be easily converted from English to Ukrainian without any complications. However, text recognition and information extraction would require a big adaptation to transfer from a Latin-based language to a Cyrillic one.

When data for a specific language are scarce, semi-supervised and self-supervised approaches can help reduce the amount of annotation required, in some cases even approaching zero-shot settings [6].

The contributions of this article are threefold. First, it frames Ukrainian handwritten text recognition as a data-bottleneck problem and documents substantial error rates of current recognition systems on Cyrillic handwriting. Second, it introduces a data-efficient generation pipeline in which glyph-composed Ukrainian phrases are refined by a prompt-controlled diffusion editing model, avoiding the need to train a dedicated handwriting generator on a large writer-labelled corpus. Third, it proposes an OCR-based self-verification loop for filtering hallucinated samples and identifies the validation steps required before this component can be used for full-corpus construction.

Related works. Key works for synthetic text generation, that can be used for neural network training, are “Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition” by Jaderberg et al. [14] and “Synthetic Data for Text Localisation in Natural Images” by Gupta et al. [11]. The main difference between the two works is the fact that in the first work, synthetic data are generated exclusively for text recognition, while in the second one generated data were used for both text detection and recognition. Additionally, in the second work, the text is attempted to be placed in real conditions, like solid-colored areas located on a single plane.

Both works use an array of approaches to make text more realistic, like adding shadows, borders, and background mixing. The second work has an accent on color selection to avoid situations when white text is placed on a white background.

Synthetic data generated in these works can be used for training text detection and recognition models. For example, in “E2E-MLT: An Unconstrained End-to-End Method for Multi-Language Scene Text” by Bušta et al. [2] use synthetic data for multi-language text recognition models, including Arabic, Bengali, Chinese, Japanese, Korean, Latin and Hindi.

In “A Multiplexed Network for End-to-End, Multilingual OCR” authors achieve improved results through more advanced approaches to text detection [13]. Their method allows for the detection of text that is difficult to enclose within a rectangular box, which is important for text in natural scenes.

A separate branch of research generates handwritten text directly. Recent methods have moved from GAN-based generators such as ScrabbleGAN [29] to latent diffusion. DiffusionPen [24] is a few-shot styled generation method based on latent diffusion models with a hybrid style extractor that combines metric learning and classification; its authors show that the generated data closely match the real handwriting distribution and improve handwritten text recognition (HTR) systems when used for training. WordStylist [28] and One-DM [27] follow the same latent-diffusion paradigm, imitating a target writer from as little as one reference sample.

Whereas most generators operate on isolated words, DiffBrush [25] targets full text-line generation: it disentangles style from content via column- and row-wise masking and adds multi-scale content supervision through line- and word-level discriminators, reporting state-of-the-art style imitation and content preservation on the IAM and CVL benchmarks. Line-level generation is

directly relevant to our pipeline, which composes and refines multi-word phrases rather than single words.

The scarcity of Ukrainian handwriting data – the very bottleneck this paper addresses - has recently begun to change. RUKOPYS [23], released in 2026 by the Ukrainian Catholic University, is described by its authors as the first large-scale open dataset for Ukrainian handwritten text recognition: 10,735 page images with region-level bounding boxes and transcriptions, spanning documents from 1919-1935 archival manuscripts to 2020-2025 school and exam sheets, distributed under the CC BY 4.0 licence. In this work it serves as an independent evaluation corpus and as evidence that real Ukrainian handwriting exhibits exactly the cross-writer variability our generator seeks to reproduce. At the same time, an evaluation dataset does not eliminate the training-data bottleneck: scalable generation is still needed to produce large volumes of diverse labelled handwriting for model development and controlled augmentation.

Most relevant to the Ukrainian setting, Ahitoliev and Berezin [26] retrain DiffusionPen on a purpose-built Ukrainian word dataset of 126,177 images from 308 writers, showing that a Latin-trained latent-diffusion generator transfers to Cyrillic without architectural changes. Our approach differs fundamentally: instead of training a diffusion generator on scarce real Ukrainian handwriting, we refine glyph-composed images with a general-purpose, prompt-controlled editing model, avoiding the need for a large writer-labelled training corpus altogether.

Testing existing solutions. To make an improvement it is important to understand the level current solutions are on right now. To test existing solutions multiple popular models were chosen: Tesseract [1], Surya [20], TrOCR-ru [19], and EasyOCR [9].

Tesseract – one of the most popular OCR engines in the world right now. It has been in development since the mid 80's and has updates coming up to this day [1]. It supports 100+ languages including Cyrillic ones like Ukrainian [20].

Surya – document OCR toolkit. It is one of the biggest open-source OCR tools, currently, it has over 13 thousand stars on GitHub and is being used by a large number of companies. It supports 90 languages including Ukrainian.[10]

TROCR-ru – Transformer-based Optical Character Recognition with Pre-trained Models. Instead of the usual approach of using CNN for image understanding and RNN for char-level text generation, this model uses Transformers architecture [16]. This particular model is fine-tuned version of the aforementioned model specifically targeted at Cyrillic text.

EasyOCR – Another popular open-source solution with more than 80 languages support and more than 24 thousand starts on Github [9].

The biggest problem in assessing existing solutions is the lack of any big datasets of handwritten Ukrainian text. That is why the creation of high-level synthetic data is a big part of achieving good results with our research. For the present benchmark, three datasets with different sampling protocols were used: the official test split of the Cyrillic Handwriting Dataset [5], the first 1,000 images of the HKR dataset [7], and 1,000 images from our early synthetic dataset.

Cyrillic Handwritten Dataset – is one of the biggest datasets available, which consists of more than 70 thousand segments of handwritten text. Unfortunately, all of those segments are written in Russian, which does not cover our needs, but should be good enough for initial assessment. In this study, the official test split was used, corresponding to approximately 5 % of the complete dataset of 73,830 handwritten text segments.

HKR dataset – Handwritten Kazakh and russian dataset. This is a big dataset that consists of more than 63 thousand sentences taken from more than 1400 filled forms. In addition to 33 Cyrillic characters used in Russian, those words also use 9 additional specific characters that are used in Kazakh. For evaluation, the first 1,000 HKR images were used.

The glyph bank is collected with a purpose-designed single-page A4 collection form. The sheet carries four solid corner fiducial squares and a column of timing marks aligned with each row of cells; together these allow a scan or photograph taken at arbitrary resolution and skew to be rectified and its cell grid localised without any learned detector. Cell outlines and instruction frames are

printed in a drop-out red, so the red channel can be discarded during preprocessing and only the pen stroke survives into the extracted image.

The sheet presents the complete Ukrainian alphabet – 33 uppercase and 33 lowercase characters, including Ґ, Є, І, Ї and the soft sign – with one empty cell per character and the target character printed immediately above it, so a completed form yields 66 labelled glyph cells, one per character. Contributors were free to return more than one form, and repeated returns are what supply the within-writer variability of the bank. Two properties of this design matter for the quality of the resulting bank. First, segmentation is a fixed geometric crop relative to the fiducials rather than a detection problem, which removes an entire class of segmentation errors. Second, the label of every cell is fixed by the printed reference character, so the bank is ground-truth labelled by construction and needs no transcription pass and no cross-referencing against any record about the writer. The full-alphabet grid also yields an approximately balanced bank, in contrast to free-text name fields, whose class distribution follows the letter frequency of personal names and over-represents a small subset of characters by two orders of magnitude.

Forms were collected by open crowdsourcing rather than from a single institutional cohort. Blank forms were distributed as a printable PDF through public call; contributors printed, completed and returned them as scans or photographs through upload form. Crowdsourced collection of handwriting corpora is established practice: the IAM [31] and CVL [32] databases were both assembled from voluntary contributors writing prescribed material, and it is this mode of collection that makes the writer population broad rather than confined to a single age band or institution

The glyph bank is not drawn from a single source, and the two components differ in what may be released. The publishable component is the crowdsourced form collection described above: 100 completed forms returned by 100 distinct contributors, each supplying one sample of every character of the alphabet. A second and considerably larger component was derived from previously digitised Ukrainian handwriting material that had already been segmented into isolated characters before this study; it was obtained under a written agreement with its holder on terms that permit processing for research purposes but not redistribution, and it is therefore not part of the release. This is the reason the per-class counts reported below are an order of magnitude larger than the form collection alone could yield, and the reason those counts are uneven across classes: the restricted component was not collected against a full-alphabet grid and carries the letter frequencies of the material it came from, whereas the form collection contributes exactly one sample per character per contributor.

Both components pass through the filtering, composition and refinement stages identically, and the counts and results reported in this paper are for the combined bank. Because each form carries exactly one sample of every character, every image contributed by the form collection traces to exactly one returned form, so the writer composition of that component is known by construction rather than inferred, which is what makes writer-disjoint splits possible within it.

Because both components can still contain hesitant, corrected, crossed-out or otherwise anomalous characters, an additional filtering stage was applied before synthetic data generation.

The results of SimCLR training can be seen in Fig. 1. Each letter forms its own cluster, while a smaller number of outliers remain scattered across the embedding space.

Before outlier filtering, the initial glyph bank contained 258,915 isolated character images representing, in the restricted component, 32 of the 33 letters of the Ukrainian alphabet; Ґ, absent from that material, is supplied by the form collection. The number of samples per class ranged from 382 to 9,962, with a mean of 8,091.09 and a median of 9,513.5 images per class. Large classes were sometimes split across several technical folders for storage, but these folders were aggregated by character label before analysis.

For glyph-bank filtering, SimCLR used a ResNet-18 encoder trained from scratch on 96 x 96 glyph images for 30 epochs with batch size 256 and learning rate 0.001. The training objective was NT-Xent with temperature 0.1; feature vectors were L2-normalized and compared with dot-product similarity. Neither dropout nor a learning-rate scheduler was used, and the random seed was fixed to 42.

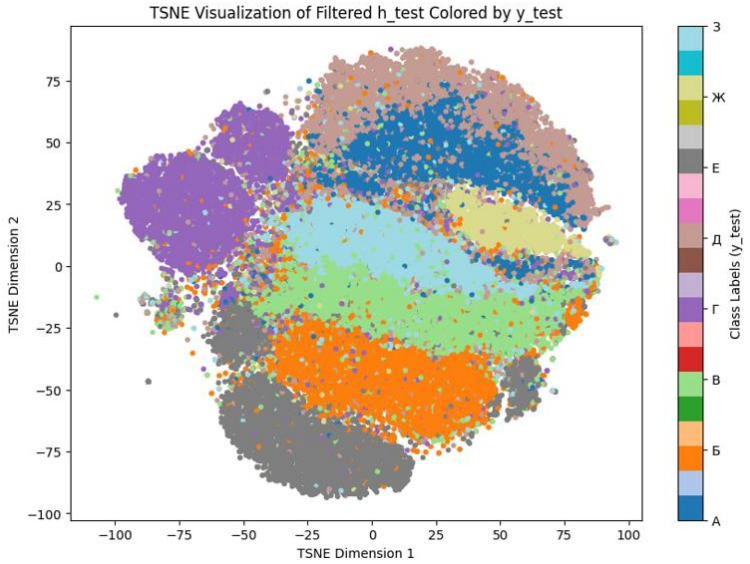


Fig. 1. **SimCLR evaluation on first 10 letters**

Source: **compiled by the authors**

After SimCLR training, embeddings were processed separately for each character class. K-means clustering [17] was used to estimate class centroids, and FAISS-based similarity search in the embedding space was used to retain the 85 % of samples closest to each centroid as the filtered subset. The 85 % retention threshold was chosen empirically during preliminary experiments as a compromise between removing anomalous samples and preserving natural intra-class variability.

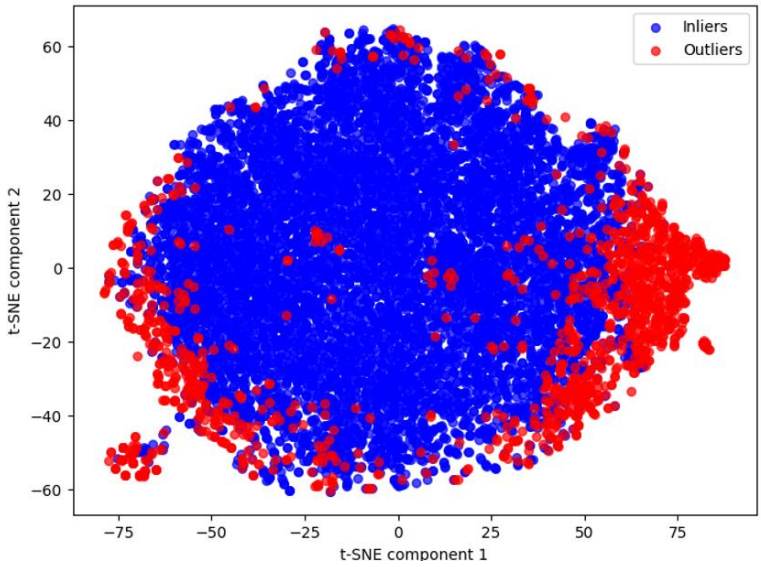


Fig. 2. **t-SNE representation of outlier detection using Faiss**

Source: **compiled by the authors**

The form carries no field for a name, contact detail, institution, place or date of birth, and the participation notice printed on it asks contributors not to write any such information anywhere on the sheet. The only identifier on a sheet is a sheet code, drawn at random from a pre-generated pool and printed as a short human-readable string; it is not derived from and never associated with any attribute of the contributor, and no linkage table between sheet codes and contributors exists at any point in the pipeline. The sheet code exists solely so that a defective print batch can be traced and withdrawn. Full-page scans are retained only until segmentation of the form collection completes and are then on access-controlled storage and never released; what leaves the pipeline is a set of isolated 96 × 96 single-character crops carrying only a character label and a sheet code.

Participation is voluntary, unremunerated and anonymous, and the notice printed on the form states the purpose of collection, the intended open release under the MIT licence, and that returning

a completed form constitutes agreement to that use. The processing basis is the contributor's consent given at the point of contribution – Art. 6(1)(a) of Regulation (EU) 2016/679 [33] and Art. 11 of the Law of Ukraine “On Personal Data Protection” [34] – read together with the research safeguards of Art. 89 GDPR. Because no linkage table was created and the released artefacts carry no direct or indirect identifier, the released bank is anonymous data in the sense of Recital 26 GDPR, as elaborated by the Article 29 Working Party opinion on anonymisation techniques [35], and therefore falls outside the scope of data-protection law. This claim is deliberately narrow. Writer identification from handwriting is a mature capability and text-independent methods reach high accuracy given a page of connected script [36]; those methods, however, draw on allographic and textural statistics accumulated over a text line – inter-letter connection patterns, slant consistency, spacing rhythm, ink-density distribution across a paragraph – none of which survives in a detached single-character crop. The composition stage further mixes glyphs from different contributors within one synthetic line, so no released line image corresponds to any real writer. We regard the residual re-identification risk as low but not nil, and for that reason the page scans from which the crops were cut are not released.

The first synthetic corpus used in the evaluation corresponds to an early synthetic dataset produced before diffusion refinement. Textual content was extracted sequentially from the UberText corpus [15] after removing emojis, punctuation, and numeric characters. The remaining words were partitioned into non-overlapping phrases of two or three words, converted to uppercase, and rendered from individual glyphs. Conventional augmentations – rotation, salt-and-pepper noise, and background variation – were then applied. Examples are shown in Fig. 3.

a b

Fig. 3. **Examples from the early synthetic dataset: a – glyph-composed phrase with conventional augmentation; b – another glyph-composed phrase with conventional augmentation**

Source: **compiled by the authors**

To evaluate models on these datasets two metrics were chosen: Character Error Rate (CER) [3] and Word Error rate (WER) [21]. There were computed via jiwer library. These two metrics are common metrics that are used for text and speech recognition. WER measures an average number of errors per word, with 0 being the perfect score. CER is doing the same but on a character level. So in our tests, we want those numbers to be as low as possible. Both metrics are based on edit operations between the reference and recognized text, including substitutions, deletions, and insertions. Because they are normalized by the reference length, CER and WER are not bounded above by 1; values greater than 1 are therefore valid when the number of required edits exceeds the length of the reference sequence.

Surya confidence filtering was calibrated in an exploratory experiment on the official test split of the Cyrillic Handwriting Dataset. One hundred thresholds uniformly spaced between 0 and 1 were evaluated, and 0.5 was selected because it gave the best trade-off on that same split. Because the threshold was tuned on the evaluation split rather than on a separate validation set, the reported Surya results should be interpreted as exploratory.

Results of the aforementioned tests can be seen in Table 1. Table 1 reports performance on three subsets: the official Cyrillic Handwriting Dataset test split, the first 1,000 HKR images, and 1,000 images from the early synthetic dataset. The early synthetic dataset refers to glyph composition followed by conventional augmentation and does not include diffusion refinement.

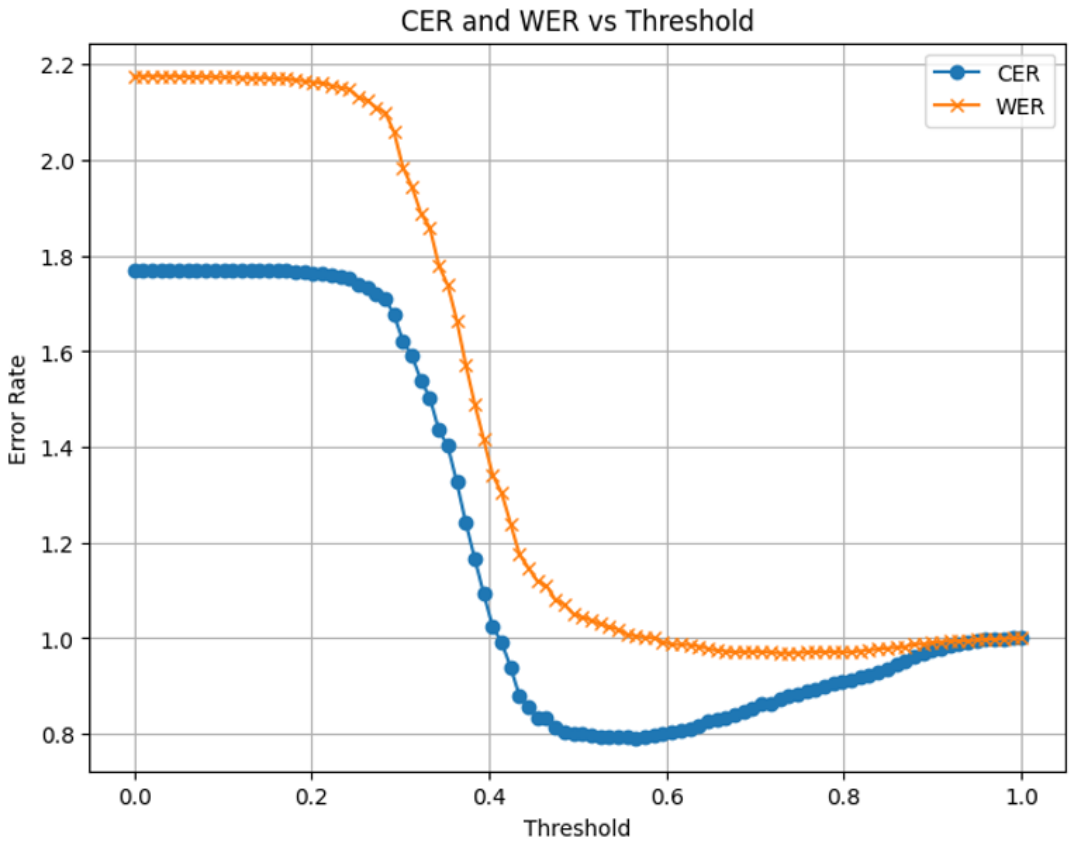


Fig. 4. **Surya CER and WER based on threshold**
Source: **compiled by the authors**

Table 1: **Test results**

Dataset	Tesseract		Surya		TROCR-ru		EasyOCR	
	CER	WER	CER	WER	CER	WER	CER	WER
Cyrillic Handwritten (test)	0.85	1.25	0.8	1.04	0.07	0.33	0.9	1.4
HKR	0.88	1.19	0.83	1.02	0.02	0.12	0.93	1.27
Early synthetic	0.7	1.27	1.69	1.97	0.46	0.94	0.75	1.12

Source: **compiled by the authors**

The three-dataset design reduces dependence on any single benchmark. The particularly low error rates of TrOCR-ru on the Cyrillic Handwritten Dataset and HKR should nevertheless be interpreted cautiously because those datasets were used during model training. On the early synthetic subset, which was not part of its training data, TrOCR-ru shows substantially higher error rates that are closer to those of the other systems. Values above 1 in Table 1 are valid under the jiwer definition described above. Each metric was computed once over the complete evaluation subset for each experiment; run-to-run variability is therefore not reported, and future work should estimate uncertainty through resampling or repeated evaluation on independently sampled subsets.

Proposed approach: diffusion-refined synthetic handwriting generation. The glyph-composition procedure described above produces images whose characters are genuine, but whose overall appearance is visibly artificial: letters from different writers are pasted next to each other without connecting strokes, a consistent baseline, or a common writing instrument. We therefore propose extending the pipeline with two additional stages – diffusion refinement and automatic verification – to obtain the seven-block architecture shown in Fig. 5.

Blocks 1-4 correspond to the steps already described: (1) automatic extraction of a per-character glyph bank from collection forms and subsequent outlier filtering; (2) denoising of the bank with SimCLR embeddings and class-centroid filtering that retains the closest 85 % of samples;

(3) sequential extraction of uppercase two- to three-word phrases from UberText after text cleaning; and (4) rendering of an early synthetic phrase image from glyphs drawn from different writers, followed by rotation, salt-and-pepper noise, and background variation.

Block 5 – diffusion refinement – sends the early synthetic image to the OpenAI Images API through the images/edits endpoint using the gpt-image-2 model [30]. No non-default API parameters were specified. Rather than using a short stylistic instruction, the method relies on a structured prompt that separates preservation constraints from realism constraints.

The prompt shown below was used for the diffusion-refinement experiments reported in this work. It instructs the model to preserve the exact transcription, character order, letterform skeleton, slant, baseline wobble, spacing, line placement, margins, background tone, speckle noise, and aspect ratio while adding stroke-level properties associated with natural handwriting, such as pressure variation, pen entry and exit traces, ink irregularity, contour micro-tremor, and occasional connecting strokes. No systematic experiment on alternative prompt variants was conducted; additional style prompting is therefore treated as future work rather than as a validated capability.

Representative prompt template used in Block 5:

```
Redraw the handwriting in the attached image so it looks like genuine pen-on-surface handwriting instead of a traced or synthetic outline.
KEEP EXACTLY AS IN THE IMAGE:
- the same text (<TARGET_TEXT>), spelling, and character order;
- the same letterform style: each Cyrillic letter keeps its shape, skeleton, and construction from the input image;
- the same slant, baseline wobble, letter-size irregularity, and word spacing;
- the same line position, margins, background colour, speckle noise, and aspect ratio.
MAKE MORE REALISTIC - only the quality of the stroke:
- real ballpoint or gel-pen dynamics with pressure-driven thickness variation;
- natural pen entry and exit strokes with tapered ends, small overshoots, and hooks where the pen lifts;
- slight ink irregularity, including faint or skipped segments and small pooling at starts, stops, and joins;
- contour micro-tremor instead of smooth vector-like edges;
- a few light connecting strokes between adjacent letters where the hand would not lift;
- crisp continuous ink, without blur, double outlines, halos, or jagged anti-aliasing artifacts.
Do not add, remove, or change any letter. Do not change the wording.
```

Block 6 – automatic verification – the verification loop is currently proposed as a quality-control stage and has not yet been applied systematically to the complete refined dataset. TrOCR-ru was selected as the initial verifier because it achieved the lowest recognition error in the benchmark above. A sample would be considered a candidate for acceptance if its character error rate did not exceed 0.1, but this threshold was used as a practical high-fidelity criterion and was not systematically optimized. In the final experimental protocol, the verifier, the downstream training model, and the final evaluator should be separated. The verification procedure should also be validated by manual inspection of subsets of both accepted and rejected samples. The acceptance rate of the loop therefore remains a future evaluation target rather than a fully measured result of the present study.

Block 7 stores the accepted pairs (image, transcription) together with style metadata – ink colour, paper type, slant – producing a labelled synthetic corpus ready for HTR training. The acceptance rate of the verification loop is itself a useful quality metric of the generator and will be reported alongside CER/WER in future experiments.

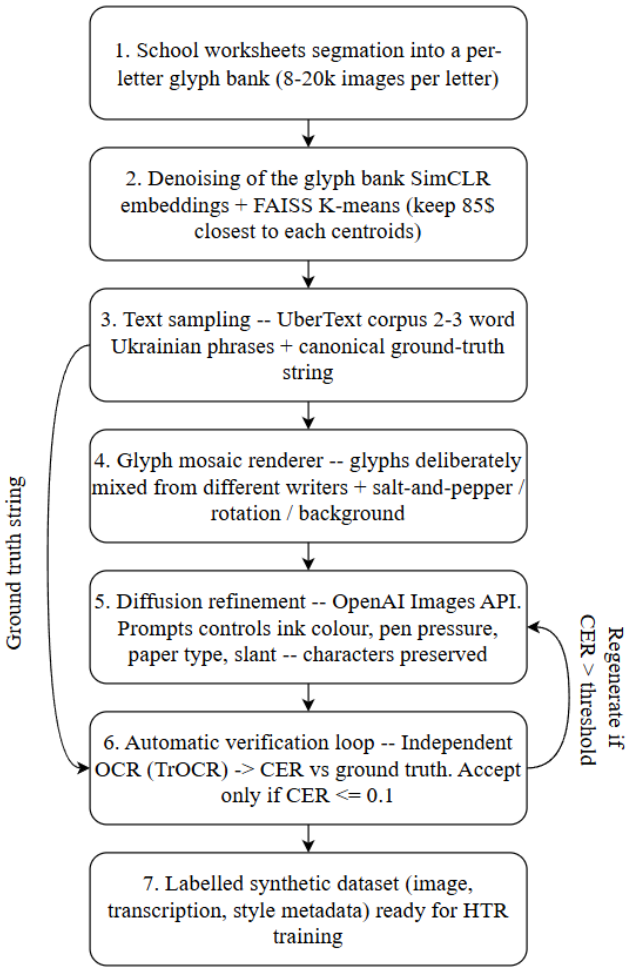


Fig. 5. **Architecture of the diffusion-refined synthetic handwriting generation pipeline**
Source: **compiled by the authors**

Potential advantages over single-style generators. First, the proposed design is intended to address harder cases than single-writer generators. Writer-conditioned methods such as DiffusionPen [24], One-DM [27] or GAN-based models [29] reproduce one handwriting style per sample. By composing glyphs from different writers within one line and then refining the result, the pipeline is designed to approximate the intra-line variability of real documents, although this advantage has not yet been tested quantitatively.

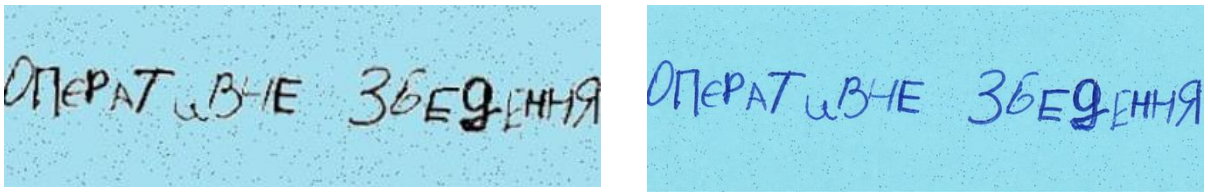
Second, the method potentially offers prompt-level control over realism without retraining. In the present study, however, only the prompt shown above was used; broader stylistic control should therefore be viewed as a future extension rather than as an experimentally validated result.

Third, the pipeline addresses the Ukrainian data bottleneck without training a dedicated generative model on a large writer-labelled corpus. Ahitoliev and Berezin [26] had to collect 126 thousand word images to retrain DiffusionPen, whereas the present approach relies on an existing image-editing model plus a filtered glyph bank. Ukrainian-specific characters present in the source bank, including є, і, and ї, are supported directly by the composition stage, including Ґ, which the full-alphabet collection form supplies.

Finally, prior literature suggests that more realistic synthetic handwriting can improve recognition systems, which motivates the proposed refinement stage. In this paper, however, the effect of the diffusion-refined corpus on downstream recognition has not yet been quantitatively evaluated.

The approach also has limitations. The editing model can hallucinate characters, and the proposed verification loop can only mitigate this risk after explicit validation. Each image is a paid API call, so large-scale generation is slower and more expensive than local sampling.

Fig. 6 provides a qualitative single example in which diffusion refinement makes an early synthetic sample appear more natural, but such an example is not sufficient to establish a systematic improvement in realism or recognition utility. Closed image-editing models may also evolve over time, which complicates reproducibility.



a b

Fig. 6. **Qualitative single example before and after diffusion refinement: a – early synthetic glyph-composed sample with conventional augmentation; b – corresponding diffusion-refined sample**

Source: **compiled by the authors**

Summary. The evaluation of existing text-recognition solutions shows that handwritten Cyrillic remains difficult for current systems and that performance varies substantially across datasets and models.

The early synthetic dataset based on glyph composition and conventional augmentation was only a first step. The diffusion-refined extension proposed here is intended to address its visual limitations by transforming these early synthetic samples into more natural handwriting, but the downstream effect of that refinement has not yet been measured quantitatively.

Future work should evaluate the refined corpus on an independent handwriting dataset and should keep verifier selection, downstream training, and final evaluation separated.

The main contribution of this work is therefore a controllable architecture for addressing the Ukrainian handwriting data bottleneck without training a dedicated writer-conditioned generator. The refined extension is demonstrated qualitatively in this paper, while its quantitative effect on recognition performance remains future work. That future evaluation should compare early synthetic and diffusion-refined data, validate the proposed verification procedure through manual audit, report acceptance behaviour, and estimate uncertainty through confidence intervals or repeated sampling.

Conflict of interests: Authors declare that they do not have conflict of interests, in particular financial, personal, copyright or any of a different nature, which could have influenced the research, as well as the results published in this articles

Financing: Research was carried out without financial support

Accessibility data: All data available in digital or graphic form in the main text manuscript.

Using artificial intelligence tools: During preparation manuscript used tool Chat GPT 5.4 of artificial intelligence (AI) for grammatical and stylistic text checks. All fragments processed by him were checked and edited by the author. AI was not used to generate text, formulas, and tables, formation scientific provisions, receipt results or formulation conclusions. Scientific the provisions, results and conclusions are my own, intellectual by the contribution of the authors

REFERENCES

1. Smith, R. “An overview of the Tesseract OCR engine”. In *Ninth International Conference on Document Analysis and Recognition (ICDAR).* 2007. p. 629–633. DOI: https://doi.org/10.1109/ICDAR.2007.4376991.

2. Bušta, M., Patel, Y. & Matas, J. “E2E-MLT – An unconstrained end-to-end method for multi-language scene text”. In *Computer Vision – ACCV 2018 Workshops. Springer*. 2019. p. 127–143. DOI: https://doi.org/10.1007/978-3-030-21074-8_11.

3. Thennal, D. K, James, J., Gopinath, D. P. & Muhammed, A. K. “Advocating character error rate for multilingual ASR evaluation”. In *Findings of the Association for Computational Linguistics*. 2025. p. 4941–4950. DOI: https://doi.org/10.18653/v1/2025.findings-naacl.277.

4. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. “A simple framework for contrastive learning of visual representations”. In *Proceedings of the 37th International Conference on Machine Learning*. 2020. p. 1597–1607. DOI: https://doi.org/10.48550/arXiv.2002.05709.

5. “Cyrillic handwriting dataset – Kaggle”. – Available from: https://www.kaggle.com/datasets/constantinwerner/cyrillic-handwriting-dataset. – [Accessed 11.11.2024].

6. Etter, D., et al. “A synthetic recipe for OCR”. In *International Conference on Document Analysis and Recognition (ICDAR)*. 2019. p. 864–869. DOI: https://doi.org/10.1109/ICDAR.2019.00143.

7. Nurseitov, D., et al. “Handwritten Kazakh and Russian (HKR) database for text recognition”. In *Multimedia Tools and Applications 80*. 2021. p. 33075–33097. DOI: https://doi.org/10.1007/s11042-021-11399-6.

8. Douze, M., et al. “The Faiss library”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2401.08281.

9. “EasyOCR – JaidedAI”. – Available from: https://github.com/JaidedAI/EasyOCR. – [Accessed 11.11.2024].

10. Paruchuri, V. & Datalab Team. “Surya: A lightweight document OCR and analysis toolkit”. 2025. – Available at: https://github.com/datalab-to/surya. –[Accessed 21.08.2025].

11. Gupta, A., Vedaldi, A. & Zisserman, A. “Synthetic data for text localisation in natural images”. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. p. 2315–2324. DOI: https://doi.org/10.1109/CVPR.2016.254.

12. Nair, A., et al. “Web archives metadata generation with GPT-4o: challenges and insights”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2411.05409.

13. Huang, J., et al. “A multiplexed network for end-to-end, multilingual OCR”. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. p. 4545–4555. DOI: https://doi.org/10.1109/CVPR46437.2021.00452.

14. Jaderberg, M., et al. “Synthetic data and artificial neural networks for natural scene text recognition”. 2014. DOI: https://doi.org/10.48550/arXiv.1406.2227.

15. Chaplynskyi, D. “Introducing UberText 2.0: A corpus of modern Ukrainian at scale”. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop*. 2023. p. 1–10. DOI: https://doi.org/10.18653/v1/2023.unlp-1.1.

16. Li, M., et al. “TrOCR: Transformer-based optical character recognition with pre-trained models”. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023; 37 (11): | 13094–13102. DOI: https://doi.org/10.1609/aaai.v37i11.26538.

17. Lloyd, S. P. “Least squares quantization in PCM”. In *IEEE Transactions on Information Theory*. 1982; 28 (2): 129–137. DOI: https://doi.org/10.1109/TIT.1982.1056489.

18. “Microsoft Azure AI Vision language support documentation”. – Available from: https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/language-support. – [Accessed 11.11.2024].

19. “raxtemur/trocr-base-ru – Hugging Face”. – Available from: https://huggingface.co/ raxtemur/trocr-base-ru. – [Accessed 11.11.2024].

20. “Tesseract OCR language data documentation”. – Available from: https://github.com/ tesseract-ocr/tessdoc. – [Accessed 11.11.2024].

21. Morris, A. C., Maier, V. & Green, P. “From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition”. In *Interspeech* . 2004. p. 2765–2768. DOI: https://doi.org/10.21437/Interspeech.2004-668.

22. Zhu, Y., Yao, C. & Bai, X. “Scene text detection and recognition: recent advances and future trends”. In *Frontiers of Computer Science*. 2016; 10 (1): 19–36. DOI: https://doi.org/10.1007/s11704-015-4488-0.

23. Voitekh, D., Zmiivskyi, V. & Molchanovskyi, O. “RUKOPYS: Ukrainian handwritten text recognition dataset”. *Ukrainian Catholic University*. 2026. – Available from: https://huggingface.co/datasets/UkrainianCatholicUniversity/rukopys. – [Accessed 20.05.2026].

24. Nikolaidou, K., Retsinas, G., Sfikas, G. & Liwicki, M. “DiffusionPen: Towards controlling the style of handwritten text generation”. In *Computer Vision – ECCV 2024, LNCS. Springer*. 2025; 15143: 417–434. DOI: https://doi.org/10.1007/978-3-031-73013-9_24.

25. Dai, G., et al. “Beyond isolated words: Diffusion brush for handwritten text-line generation”. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025. p. 19054–19064. DOI: https://doi.org/10.1109/ICCV51701.2025.01771.

26. Ahitoliev, A. & Berezin, P. “Diffusion-based Ukrainian handwritten text generation with cross-domain style transfer”. *arXiv*. 2026. DOI: 10.48550/arXiv.2605.27487.

27. Dai, G., et al. “One-DM: One-shot diffusion mimicker for handwritten text generation”. In *Computer Vision – ECCV. Springer*. 2024. p. 410–427. DOI: https://doi.org/10.1007/978-3-031-73636-0_24.

28. Nikolaidou, K., et al. “WordStylist: Styled verbatim handwritten text generation with latent diffusion models”. In *Document Analysis and Recognition – ICDAR*. LNCS. *Springer*. 2023; 14188: 384–401. DOI: https://doi.org/10.1007/978-3-031-41679-8_22.

29. Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S. & Litman, R. “ScrabbleGAN: Semi-supervised varying length handwritten text generation”. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. p. 4324–4333. DOI: https://doi.org/10.1109/CVPR42600.2020.00438.

30. “OpenAI Images API – image editing documentation”. – Available from: https://platform.openai.com/docs/guides/images. – [Accessed 20.05.2026].

31. Marti, U.-V. & Bunke, H. “The IAM-database: an English sentence database for offline handwriting recognition”. In *International Journal on Document Analysis and Recognition*. 2002; 5: 39–46. DOI: https://doi.org/10.1007/s100320200071.

32. Kleber, F., Fiel, S., Diem, M. & Sablatnig, R. “CVL-DataBase: An off-line database for writer retrieval, writer identification and word spotting”. In *12th International Conference on Document Analysis and Recognition (ICDAR)*. 2013. p. 560–564. DOI: https://doi.org/10.1109/ICDAR.2013.117.

33. “European Parliament and Council of the European Union. Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)”. *Official Journal of the European Union*. 2016; L 119: 1–88. Recital 26; Arts. 6, 89. – Available from: https://data.europa.eu/eli/reg/2016/679/oj. – [Accessed 15.05.2026].

34. “Verkhovna Rada of Ukraine”. *Law of Ukraine No. 2297-VI of 1 June 2010 “On Personal Data Protection.”* Arts. 5, 11. – Available from: https://zakon.rada.gov.ua/laws/show/2297-17?lang=en. – [Accessed 15.05.2026].

35. “Article 29 Data Protection Working Party”. *Opinion 05/2014 on Anonymisation Techniques*. WP216. Adopted 10 April 2014. Brussels: European Commission. 2014. – Available from: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf/ – [Accessed 15.05.2026].

36. Bulacu, M. & Schomaker, L. “Text-independent writer identification and verification using textural and allographic features”. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2007; 29 (4): 701–717. DOI: https://doi.org/10.1109/TPAMI.2007.1009.

Received 22.07.2026
Received after revision 27.08.2026
Accepted 03.09.2026

DOI: https://doi.org/10.15276/ict.03.2026.09
УДК 004.855.91:514.112

Від ізольованих гліфів до реалістичного рукопису: генерація синтетичних даних із дифузійним уточненням для розпізнавання українського кириличного тексту

Березін Павло Романович
ORCID: https://orcid.org/0000-0003-1869-5050; pavlo.berezin@gmail.com
Національний університет «Києво-Могилянська академія», вул. Сковороди 2, Київ 04070, Україна

АНОТАЦІЯ

Розпізнавання рукописної кирилиці, і зокрема української, суттєво відстає від латиниці — насамперед через історичну відсутність великих анотованих датасетів рукопису. У цій роботі брак навчальних даних розглянуто як центральну проблему, яку розв'язано через генерацію синтетичних даних. Спершу протестовано поширені системи оптичного розпізнавання символів (Tesseract, Surya, TrOCR, EasyOCR) на кириличному рукописі й підтверджено, що похибки на рівні символів і слів залишаються суттєво вищими, ніж для англійської, що мотивує підхід, орієнтований на дані. Головний внесок полягає у створенні наскрізного конвеєра генерації синтетичного українського рукопису. Він формує банк ізольованих гліфів зі стандартизованих конкурсних бланків, відфільтровує аномальні зразки у вивченому просторі ознак, добирає очищений фразовий вміст із сучасного українського корпусу, компонує багатослівні рядки, застосовує звичайні візуальні аугментації та уточнює результати за допомогою керованої підказкою універсальної моделі редагування зображень, щоб наблизити вигляд до природного письма. Робота також пропонує автоматичний етап верифікації, покликаний зменшити ризик потрапляння до майбутнього навчального корпусу візуально правдоподібних, але текстово некоректних зразків. Архітектура дає змогу поєднувати в одному рядку гліфи різних авторів, зберігаючи цільову транскрипцію та керовано уточнюючи вигляд письма. У поточному дослідженні дифузійно уточнене розширення показано якісно, а не через повну кількісну перевірку його впливу на подальше розпізнавання. Отже, у статті подано керовану архітектуру синтезу українського рукопису, окреслено її обмеження та визначено протокол валідації, потрібний для встановлення її цінності для розпізнавання.

Ключові слова: генерація синтетичних даних; генерація рукописного тексту; дифузійні моделі; редагування зображень за текстом; кириличне розпізнавання символів; український рукопис; розпізнавання рукописного тексту; аугментація даних

ABOUT THE AUTHOR

Pavlo Berezin - Graduate Student, Department of Computer Science. National University of Kyiv Mohyla Academy, 2, Skovorody Vul. Kyiv 04070, Ukraine
ORCID: https://orcid.org/0000-0003-1869-5050; pavlo.berezin@gmail.com
Research field: computer vision

Березін Павло Романович - Аспірант кафедри Комп'ютерних наук. Національний університет «Києво-Могилянська академія», вул. Сковороди 2. Київ 04070, Україна
Галузь досліджень: комп'ютерний зір