

Artificial Intelligence-Driven Text-to-Tactile Graphics Generation for Visual Impaired People

Yehor Dzhurynskiy¹, Volodymyr Mayik² and Lyudmyla Mayik²

¹ Ukrainian Academy of Printing, 19, Pid Holoskom Str., Lviv, 79020, Ukraine

² Lviv Polytechnic National University, 28a, Stepan Bandera Str., Lviv, 79013, Ukraine

Abstract

This research presents the development of a text-conditional tactile graphics generation model using the Bidirectional and Auto-Regressive Transformer (BART) and Vector Quantized Variational Auto-Encoder (VQ-VAE). The model leverages a modified organization of the latent space, divided into two independent components: textual and graphic. The study addresses the challenge of the limited availability of tactile graphics samples by expanding the training dataset with custom samples, enhancing the model's capability to convert textual information into graphical representations. The proposed method improves the creation of tactile graphics for visually impaired individuals, offering increased variability, controllability, and quality in synthesized tactile graphics. This advancement enhances both the technical and economic aspects of the production process for inclusive educational materials.

Keywords ¹

Artificial intelligence, tactile graphics, visual impairment, natural language processing, model, machine learning

1. Introduction

The dynamics of modern inclusive society development emphasize the need to integrate people with visual impairments into active social life. The problem of socializing individuals with visual impairments involves various aspects that complicate their education, training, and full participation in society [1]. Specifically, people with visual impairments have limited access to information, as many materials are produced only in the usual printed or digital formats. This issue is further exacerbated by the increasing prevalence of information in graphic form, designed for more effective perception by readers. The aforementioned problems hinder the ability of individuals with visual impairments to receive quality education and professional development [2, 3, 4, 5, 7].


An analysis [6] of the activities of publishing and printing industry enterprises that produce educational and methodological literature (textbooks, manuals, etc.) for people with visual impairments revealed problems related to the creation or adaptation of images and illustrative materials, which are particularly crucial for this type of publication. When creating or adapting graphic materials, enterprises encounter the following issues: an insufficient number of trained specialists with specific competencies related to the technical implementation of tactile graphics; additional time and financial costs for training specialists; and the high labor intensity and cost of the process of creating or adapting tactile graphics. Consequently, the production issues surrounding tactile graphics remain one of the primary factors contributing to the low level of access to graphical information for people with visual impairments.

ProfIT AI 2024: 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25–27, 2024, Cambridge, MA, USA

✉ y.a.dzhurynskiy@gmail.com (Y. Dzhurynskiy); vol.mayik.2015@gmail.com (V. Mayik); ludmyla.maik@gmail.com (L. Mayik)

ORCID iD 0000-0002-6650-2703 (V. Mayik); 0000-0001-8552-0942 (L. Mayik)

© 2024 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Work

Scientists are working to solve the problem of producing tactile graphics by developing models for the automatic generation of tactile images [8, 9, 10, 11, 12, 13]. The task of most existing models is to transform the content of a photo image into a tactile one.

Models [8, 9, 10, 11] that attempt to directly convert the content of an image into a tactile format usually utilize computer vision and have the following disadvantages: they violate the requirements for tactile graphics [14, 15]; they display redundant elements of the image that are difficult to read and interfere with the overall interpretation of the graphic material.

In models [12, 13] whose principle of operation involves the detection and subsequent recognition of individual image elements, replacing these elements with their tactile representations from a limited sample, there is no variability in the synthesized image samples (it is impossible to synthesize new samples, and the attractiveness of the synthesized image for people with visual impairments decreases). Despite the mentioned drawback, it should be noted that the method effectively conveys the content of the original photo at a high level in compliance with the requirements for tactile graphics.

Additionally, such methods require supplementary source graphic information (e.g., photographs), the search for or creation of which slows down the process of preparing material for the production of tactile images.

The development of information technologies, particularly in the field of deep machine learning, has opened new opportunities for addressing the aforementioned problems. Recently, significant advancements have been demonstrated by information technologies based on artificial intelligence [16, 17, 18], which enable the generation of images based on user text prompts. However, according to the analysis [19, 20], confirmed by a series of experiments, the information technologies built upon these mathematical models have proven ineffective for creating tactile graphics. Despite this, the concept of text-guided image generation was chosen as the foundation for this work.

3. Text-conditional tactile graphics generation model

The text-conditional tactile graphics generation model is built upon the Bidirectional and Auto-Regressive Transformer (BART) [21] and Vector Quantized Variational Auto-Encoder (VQ-VAE) [22]. The subject of its modeling is the process of converting text information into graphic information. To do this, the embedded space of the transformer, which was formed during language modeling on pretraining task, was divided into two independent embedded spaces: text and graphics, instead of a shared one. At the same time, the parameters of the graphic embedded space were adjusted so that the dimension of the embedded space was equal to the size of the "codebook" [22], and the dimensionality of the vectors of the graphic embedded space was equal to the dimensionality of the latent space vectors of the variational image synthesis model. The parameters of the text embedded space remained the same as during language modeling.

Before obtaining text tokens using the BPE [22, 23] tokenization model, the original text components are normalized by bringing them to a uniform format (uppercase letters were converted to lowercase letters).

Formally, the process of converting text tokens into graphic tokens using a text-conditional tactile graphics generation model is described in successive stages.

The first step is to generate a bounded sequence of text tokens based on a text prompt: $\bar{t} = \{t_i \in V\}_{i=1}^{Seq_{max,t}}$, \bar{t} where is a sequence of text tokens of dimension $Seq_{max,t} = 64$; V is a dictionary of tokens. If the size of the generated sequence of text tokens exceeds the value, its size is reduced to the maximum value, discarding the excess tokens. If the size of the generated sequence of text tokens is smaller than the value, its size is increased to the maximum value by adding utility tokens (PAD) that do not affect the simulation result.

In the next step, the text tokens that form the sequence \bar{t} are mapped to the text embedded space vectors e_k , forming a subset of it:

$$\bar{e}^t = \{e_k^t \in E^t | k = t_i \in \bar{t}\}_{i=1}^{Seq_{max,t}}; \bar{e}^t \subseteq E^t, \quad (1)$$

where \bar{t} is the sequence of text tokens; E^t is a text embedded space; e_k^t are elements of the text embedded space. Elements \bar{e}^t reflect the semantic meaning of text tokens in the embedded space.

Next, the vectors of the text embedded space E^t are transformed by the transformer's bidirectional encoder, which is formed from several layers, forming hidden states \bar{h}^t . The bidirectionality of the encoder means that it analyzes the full context of an individual vector of the embedded space, considering both the previous and the following elements of the sequence:

$$\bar{h}^t = Encode(\bar{e}^t); \bar{h}^t \subseteq E^t, \quad (2)$$

where \bar{h}^t is the hidden state of the encoder; $Encode(\cdot)$ is the transformer's encoding operation defined within [21].

The hidden state of the encoder \bar{h}^t is then converted by linear layers and a nonlinear activation function to the hidden state of the decoder (i.e., graphic information), forming a subset of the graphic embedded space E^g :

$$\bar{h}^g = Linear_2 \circ ReLU \circ Linear_1(\bar{h}^t), \quad (3)$$

where $\bar{h}^t \subseteq E^t$ is the hidden state of the encoder; $\bar{h}^g \subseteq E^g$ is the hidden state of the decoder; $Linear_i$ is a linear layer; $ReLU \stackrel{\text{def}}{=} \max(0, x)$ is a non-linear layer, activation function.

At the next stage, an autoregressive [25, 26] transformer decoder is used. This means that the decoder generates one graphics token per iteration, considering the context of the previously generated graphics tokens. Thus, during the decoding process, the model performs calculations based on the hidden state \bar{h}^t and pre-generated elements of the vector sequence of the graphic embedded space e_k^g , or \bar{e}^g :

$$e_i^g = Decode(\bar{h}^t, e_1^g, e_2^g, \dots, e_{i-1}^g); i \leq d_z, \quad (4)$$

where e_i^g is the i -th element of the vector sequence of the graphic embedded space E^g ; e_j^g ; $j < i$ are previously generated vectors of the graphic embedded space; \bar{h}^g is the hidden state of the decoder; d_z is the size of the final sequence $\bar{e}^g \subseteq E^g$; $Decode(\cdot)$ is the transformer's decoding operation defined within [21].

Decoding occurs in an iterative manner until the sequence \bar{e}^g size is equal to d_z (i.e., the size of the latent space vector of the VQ-VAE model).

Once the decoding is complete, the resulting sequence of vectors of the graphics embedded space \bar{e}^g is converted by a linear layer and *Softmax* function into a sequence of probability distributions from which the element with the highest probability is selected, determining the selected graphics token:

$$\bar{g} = \{g_i\}_{i=1}^{d_z}; g_i = arg \max (Softmax \circ Linear(e_i^g)), \quad (5)$$

where \bar{g} is the generated sequence of graphic tokens of size d_z ; $e_i^g \in \bar{e}^g$ is an element of the vector sequence of the graphic embedded space E^g .

In the next step, on the basis of graphic tokens (5), a sequence of latent quantized vectors is formed z_q , which is defined by the formula (6). Each graphic token: g_i ; $1 \leq g_i \leq K$; $i = 1..d_z$, is the positional number of the quantized vector in the "codebook" of the VQ-VAE model:

$$z_q = \{e_k \in Z | k = g_i \in \bar{g}\}_{i=1}^{d_z}; z_q \subseteq Z, \quad (6)$$

where Z is the set of latent quantized vectors, or "codebook"; $z_q \subseteq Z$ is a sequence of latent quantized vectors; $g_i \in \bar{g}$ is a graphic token; d_z is the size of the sequence of latent quantized vectors.

The final step is the synthesis of tactile graphics using a sequence-based variational image synthesis model decoder (6):

$$Y = ImDecode(z_q), \quad (7)$$

where z_q is the sequence of latent quantized vectors; $ImDecode(\cdot)$ is an image decoding operation based on latent representation defined within [22]; Y is a generated tactile image. The diagram of the text-conditional tactile graphics generation model is shown in Figure 1.

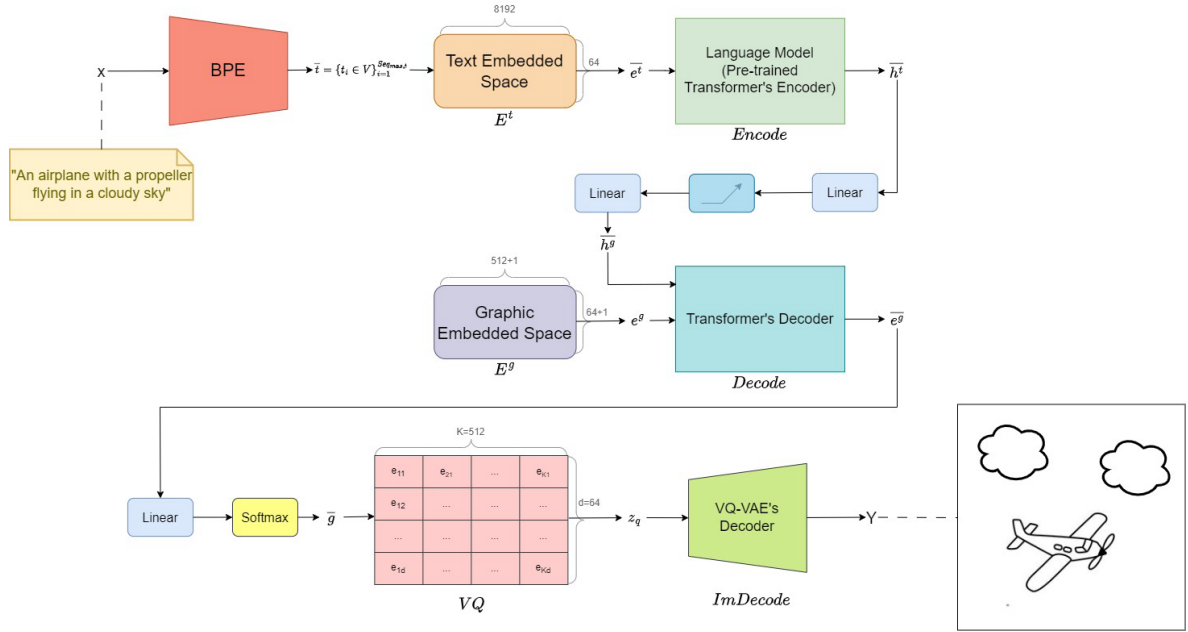


Figure 1: Structural and functional diagram of the text-conditional tactile graphics generation model

4. Experiment

In this experiment, the proposed model was trained using the parameters presented in Tables 1 and 2 for the BART and VQ-VAE models, respectively. It is important to note that the size of the decoder's dictionary and the length of the sequence are each increased by one unit compared to the original values. This adjustment is necessary to introduce an additional image service token (i.e., SOS token), which is added at the beginning of the sequence to facilitate autoregressive image generation.

Table 1
BART's parameters

Parameter	Encoder's value	Decoder's value
Dictionary size	8192	513
Sequence size	64	65
Number of layers	3	3
Layer dimension	512	512
FFN dimension	1024	1024
Number of attention heads	8	8

The language modeling has been done using the BrUK corpus [27] consisting of Ukrainian texts from different sources. Unlike textual datasets (i.e., corpora), which are widely accessible, tactile graphics samples are much less common. A significant obstacle in modeling tactile graphics generation using machine learning is the insufficient number of publicly available samples, as the tactile graphics production industry is less prevalent compared to the traditional one.

Nevertheless, a collection of plant and animal images stored in the APH Tactile Graphics Library [28] was chosen as the original set of images for the model to learn to reproduce. Additionally, the training dataset was expanded with 41 custom tactile image samples, increasing the total number of samples to 179. The custom samples are formed from simple images of animals and were used at one of the enterprises of Ukraine, which provides preschool education for children with visual impairments.

Table 2
VQ-VAE’s parameters

Parameter	Value
Image dimension	256 × 256 × 1
“Codebook” size	512
Latent vectors size	16
Number of hidden layers	5
Hidden layers dimension	16

The results of the experiment include samples of generated tactile graphics images based on various types of text prompts, such as monosyllabic prompts, prompts with numerals, and prompts with epithets. These samples are presented in Figure 2.

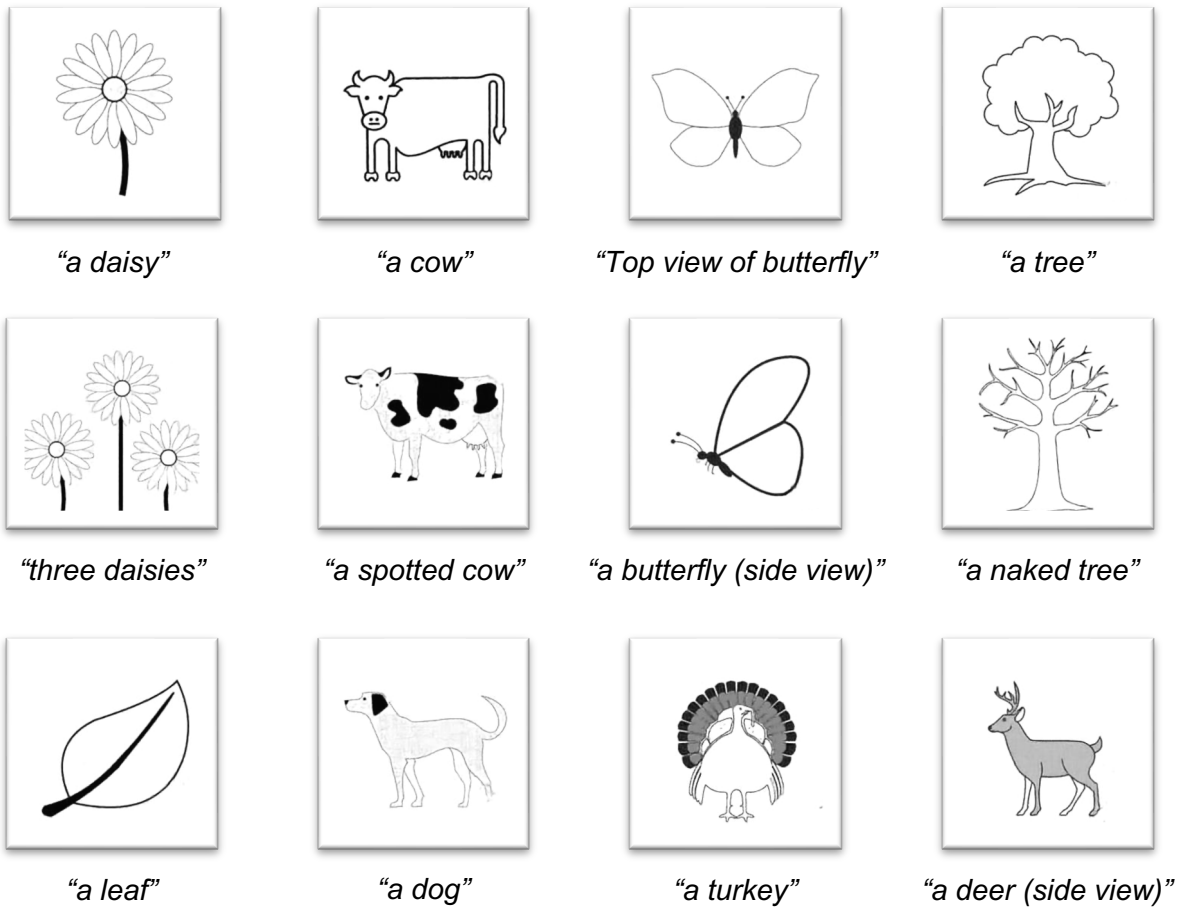


Figure 2: Samples of generated images determined by a text prompt given below the corresponding image

The model's performance was evaluated separately for each component: BART and VQ-VAE. The results of this evaluation are presented in Table 3. The Cross-Entropy metric reflects how well the model converts text prompts into appropriate graphic tokens, and Perplexity represents the uncertainty in the model's predictions. Lower values indicate better performance, meaning the model is more confident in its generation process. For tactile graphics, FID measures how similar the generated tactile images are to real ones in the latent space of the model. A lower FID score indicates that the generated tactile graphics are closer to real tactile images in terms of visual and tactile features.

Additionally, the overall performance of the model was evaluated using the CLIP Score metric [29], which reflects the model's capability in converting textual information into graphical information. The average CLIP Score of the developed model is 23,7.

Table 3
Evaluation results

Component	Metric	Value
BART	Cross-Entropy	5,709
	Perplexity	301,662
VQ-VAE	MSE (image space)	0,0144
	MSE (latent space)	0,0058
	FID	0,242

5. Limitations

The current dataset used for training includes relatively simple images (e.g., animals, plants, basic objects). One limitation of the model is its potential difficulty in scaling to more complex images, such as those with intricate details (e.g., architectural blueprints, detailed scientific diagrams). The model's ability to capture fine details may be limited by the size of the latent space and the number of hidden layers used in the VQ-VAE model. Complex tactile graphics might require a more fine-grained representation, which could lead to inefficiencies or inaccuracies in generation if the model architecture remains unchanged.

Besides, while the model performs well on simpler prompts (e.g., "a cow," "a tree"), more complex and nuanced prompts (e.g., "a group of children playing soccer with a spotted ball") might pose challenges. This is because the Transformer's encoding of textual information becomes more demanding as the semantic richness and length of the prompt increase. The model may struggle to disentangle and appropriately represent all components of a complex scene in tactile graphics form, leading to loss of information or oversimplification.

Regarding computational requirements, the training process of the proposed model, which integrates both the BART Transformer and the VQ-VAE, requires significant computational resources. Due to the autoregressive nature of the model and the need to process both textual and graphical latent spaces, training is computationally expensive. It requires powerful GPUs or TPUs, large memory capacity, and extended training time, particularly as the dataset grows. This makes scaling to larger datasets or higher-dimensional image outputs challenging without access to advanced computing infrastructure.

One of the key ethical concerns in the development of tactile graphics is ensuring that the generated images do not misrepresent the information. For visually impaired users, the tactile graphic is a primary means of understanding visual content, and any distortion or inaccuracy could lead to misunderstandings. For example, if a generated tactile graphic oversimplifies or omits important details, users might receive an incomplete or misleading representation of the intended information. To mitigate this risk, it's important to validate the model outputs rigorously against established standards for tactile graphics and seek feedback from visually impaired users to ensure that the tactile representations are both accurate and understandable.

6. Conclusion

As a result of this research, a text-conditional tactile graphics generation model was developed using BART and VQ-VAE. The model employs a modified organization of the latent space, divided into two independent components: textual and graphic.

The method of creating tactile graphics for publications aimed at individuals with visual impairments has been improved. This enhancement increases the variability, controllability, and quality of synthesized tactile graphics, thereby improving the technical and economic aspects of the production process.

This technology can bridge the gap in access to educational materials, allowing visually impaired individuals to better engage with subjects that rely heavily on visual content, such as science, mathematics, and geography. The availability of automated tactile graphics can facilitate greater independence in learning and enhance participation in inclusive classrooms and professional environments.

An important direction of further research is to increase the size and diversity of the training sample to improve the general ability of the model to generalize and ensure its stable operation in various scenarios.

References

- [1] GBD 2019 Blindness and Vision Impairment Collaborators; Vision Loss Expert Group of the Global Burden of Disease Study. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study, *Lancet Glob Health* (2021) e130-e143. doi: 10.1016/S2214-109X(20)30425-3.
- [2] P. Ackland, Serge Resnikoff, R. Bourne, World blindness and visual impairment: Despite many successes, the problem is growing, *Community Eye Health Journal* (2018) 71–73. PMID: 29483748.
- [3] K. Zebehazy and A. Wilton, "Graphic Reading Performance of Students with Visual Impairments and Its Implication for Instruction and Assessment," *Journal of Visual Impairment & Blindness*, vol. 115, pp. 215-227, 2021.
- [4] M. Mukhiddinov and K. Soon-Young, "A Systematic Literature Review on the Automatic Creation of Tactile Graphics for the Blind and Visually Impaired," 2021.
- [5] F. Bara, "The Effect of Tactile Illustrations on Comprehension of Storybooks by Three Children with Visual Impairments: An Exploratory Study," *Journal of Visual Impairment & Blindness*, vol. 112, pp. 759-765, 2018.
- [6] V. Mayik, T. Dudok, L. Mayik, N. Lotoshynska, I. Izonin, J. Kusmierczyk, An Approach Towards Vacuum Forming Process Using PostScript for Making Braille, in: *Advances in Computer Science for Engineering and Manufacturing*, Springer International Publishing, 2022, pp. 38–48. doi: 10.1007/978-3-031-03877-8_4.
- [7] Y. Dzhurynskyi and V. Mayik, "Analysis of the process of preparing illustrations for inclusive literature," *Qualilogy of the book*, vol. 41, pp. 7-15, 2022.
- [8] T. Way and K. Barner, "Towards Automatic Generation of Tactile Graphics," *Rehabilitation Engineering and Assistive Technology Society of North America*, pp. 161-163, 1996.
- [9] T. Way and K. Barner, "Automatic visual to tactile translation - Part I: Human factors, access methods, and image manipulation," *IEEE Transactions on Rehabilitation Engineering*, pp. 81-94, 1997.
- [10] T. Way and K. Barner, "Automatic visual to tactile translation. II. Evaluation of the TACTile image creation system," *IEEE Transactions on Rehabilitation Engineering*, pp. 95-105, 1997.
- [11] T. Ferro and D. Pawluk, "Automatic image conversion to tactile graphic," in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, Bellevue Washington, 2013.

- [12] K. Pakénaité, P. Nedelev, E. Kamperou, M. Proulx and P. Hall, "Communicating Photograph Content Through Tactile Images to People With Visual Impairments," *Frontiers in Computer Science*, vol. 3, 2022.
- [13] K. Pakenaite, E. Kamperou, M. J. Proulx, A. Sharma and P. Hall, "Pic2Tac: Creating Accessible Tactile Images using Semantic Information from Photographs," in *Proceedings of the Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction*, Cork, 2024.
- [14] Polish association of the blind, "Instructions for creating and adapting illustrations and typhlographic materials for blind students," 2016.
- [15] Braille Authority of North America & Canadian Braille Authority, "Guidelines and Standards for Tactile Graphics," 2022. [Online]. Available: <https://www.brailleauthority.org/guidelines-and-standards-tactile-graphics>. [Accessed 20 April 2024].
- [16] J. Oppenlaender, "The Creativity of Text-to-Image Generation," in *Academic Mindtrek '22: Proceedings of the 25th International Academic Mindtrek Conference*, New York, 2022.
- [17] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *ArXiv*, vol. abs/2204.06125, 2022.
- [18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, Louisiana, 2022.
- [19] Y. Dzhurynskyi and V. Mayik, "Preparation of illustrations for inclusive literature using artificial intelligence models of image synthesis from text," *Proceedings*, vol. 66, no. 1, pp. 155-163, 2023.
- [20] Y. Dzhurynskyi, "Generation of illustrations for inclusive literature using Midjourney artificial intelligence model," in «Scientific method: reality and future trends of researching»: collection of scientific papers «SCIENTIA» with Proceedings of the II International Scientific and Theoretical Conference, Zagreb, 2023.
- [21] Y. Yingchen, Z. Fangneng, W. Rongliang, P. Jianxiong, C. Kaiwen, L. Shijian, M. Feiying, X. Xuansong and M. Chunyan, "Diverse Image Inpainting with Bidirectional and Autoregressive Transformers," *arXiv*, vol. abs/2104.12335, 2021.
- [22] A. van den Oord, O. Vinyals and K. Kavukcuoglu, "Neural Discrete Representation Learning," *CoRR*, vol. abs/1711.00937, 2017.
- [23] Kulchytska, Kh., Semeniv, M., Kovalskyi, B., Pysanchyn, N., Selmenska, Z.: Influence of Hadamard matrices canonicity on image processing. In: Hu, Z., Petoukhov, S., Yanovsky, F., He, M. (eds.) *ISEM '21, LNCS*, vol. 463, pp. 329–338. Springer, Cham (2022) doi:10.1007/978-3-031-03877-8_29
- [24] V. Zouhar, C. Meister, J. Luis Gastaldi, L. Du, T. Vieira, M. Sachan and R. Cotterell, "A Formal Perspective on Byte-Pair Encoding," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, 2023.
- [25] M. Dalal, A. C. Li and R. Taori, "Autoregressive Models: What Are They Good For?," *CoRR*, vol. abs/1910.07737, 2019.
- [26] A. Graves, "Generating Sequences With Recurrent Neural Networks," *CoRR*, vol. abs/1308.0850, 2013.
- [27] A. Rysin, "LanguageTool API NLP UK," 2022. [Online]. Available: https://github.com/brown-uk/nlp_uk. [Accessed 21 April 2024].
- [28] American Printing House, "Tactile Graphic Image Library," [Online]. Available: <https://imagelibrary.aph.org/portals/aphb/#page/welcome>. [Accessed 21 April 2024].
- [29] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras and Y. Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," *CoRR*, vol. abs/2104.08718, 2021.