Enhancing adapted print publication accessibility via text-to-image synthesis

Rostyslav Zatserkovnyi[1,*,†], Petro Kutsyk[1,†], Roksoliana Zatserkovna[2,†], Volodymyr Maik[2,†] and Peter T. Popov[3,†]

[1] *Lviv University of Trade and Economics, 10 Tuhan-Baranovskyi Str., 79008 Lviv, Ukraine*
[2] *Ukrainian Academy of Printing, 19 Pid Goloskom Str., 79020 Lviv, Ukraine*
[3] *City University of London, Northampton Square, London, EC1V 0HB, United Kingdom*

Abstract

One of the most pressing concerns in the field of adapted printed publications – that is, publications with additional supporting features to make them easily accessible by a wide variety of audiences – is preparing illustrations that can clearly convey visual information to the viewer. These illustrations need to be created while accounting for the needs of a diverse inclusive audience, whose requirements may be affected by disabilities such as visual impairment. Currently, there is a limited number of illustrators who can appropriately produce a large number of illustrations which satisfy these requirements; therefore, illustrating print publications is a time-consuming and expensive process for non-profit organizations which are responsible for their production.

This article proposes a method for enhancing illustrations within print publications, given the source file (such as a PDF file) of a print publication. Based on modern text-to-image generators, this method extracts all illustrations from a print publication; converts them into textual prompts for a modern text-to-image generator; and finally, produces a series of adapted alternatives for each of the chosen illustrations based on the textual prompts. This allows publishers to obtain accessible illustrations for their publication in a manner of minutes, speeding up the adaptation process and enhancing its accessibility.

Keywords

Accessible publishing, artificial intelligence, image synthesis, information technologies.

1. Introduction

In recent years, assistive technologies have enabled diverse audiences of readers to freely access written information, allowing them to work, study, and participate in civil society. One of the most essential tools for this category of readers is accessible print publications. While people with disabilities, such as visual impairment, often rely on e-readers; these may not always be available or preferred. A recent study suggests that 65% of surveyed Americans have recently

ICyberPhyS-2024: 1st International Workshop on Intelligent & CyberPhysical Systems, June 28, 2024, Khmelnytskyi, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

zatserkovnyi.rostyslav@gmail.com (R. Zatserkovnyi); kutsykpetro@gmail.com (P. Kutsyk); zatserkovna.r@gmail.com (R. Zatserkovna); vol_maik@meta.ua (V. Maik); p.t.popov@city.ac.uk (P. Popov)

0000-0001-6991-2866 (R. Zatserkovnyi); 0000-0001-5795-9704 (P. Kutsyk); 0000-0003-1011-053X (R. Zatserkovna); 0000-0002-6650-2703 (V. Maik); 0000-0002-3434-5272 (P. Kutsyk)

read a print book, while only 30% have read an e-book, indicating that print books still remain the most popular format for general readers [1]. In a classroom setting in particular, print books provide unique benefits such as ease of use and improved notetaking [2]. Thus, by providing access to appropriately adapted books, educational institutions as well as independent organizations can make sure that visually impaired people can succeed.

Incorrectly adapted printed books can pose various issues for readers with visual impairment, such as small text size, poor font styling, insufficient contrast, graphics and tables as well as other design factors. Unlike e-books, where these parameters can be custom-tailored to the needs of a particular reader, printed publications are inflexible in their design, exacerbating these issues [3]. This article focuses on improving the accessibility of illustrations in particular, proposing a multi-step method to enhance the accessibility of all illustrations within a given publication:

1. Extract all illustrations from the source file of a future print publication.
2. Convert all illustrations into prompts for a text-to-image synthesizer.
3. Synthesize new, adapted illustrations based on modified versions of these prompts.
4. Review adapted illustrations and re-introduce them into the source file.

2. Related works

Existing studies in the field most often focus on e-book accessibility. Since e-book formatting is not rigid and preset by the publisher, but rather determined by the e-reader, the focus on e-book accessibility lies primarily in correctly marking important parts of the publication (such as chapters or specific phrases which link to footnotes), as well as developing robust, flexible e-reader software that allows users to adjust fonts, text sizes, as well as other parameters [4].

However, print books still have notable advantages, especially in the field of education. Research suggests that readers tend to understand text slightly better when it’s printed rather than viewed on-screen [5], and one study suggests the haptic feedback of a touch screen (or PC monitor) is different than that of a paper book, providing a less immersive experience [6].

When it comes to adapting illustrations within print books, existing research focuses on multimodal illustrations – for instance, tactile illustrations, which combine visual illustrations with tactile Braille overlays, which is useful for readers with legal blindness [7, 8]. Still, there remains the issue of creating effective adapted illustrations before their tactile component is factored in. In recent years, AI algorithms, such as machine learning algorithms and neural networks, have been trained to produce a variety of media content – and creating such illustrations from scratch is one potential application of these generative methods [9].

3. Proposed methodology

3.1. Converting images to text prompts

In order to synthesize new adapted illustrations for a print publication, our method must first obtain the inputs – often known as “prompts” – for a well-known image generator such as Stable Diffusion or Midjourney. Although in technical publications, illustrations often come with captions describing the image, these captions are not always present, and often offer brief interpretations which do not capture the full nuance of the captioned image. Thus, the images must be converted to textual prompts using an AI model known as an *image captioner.* Acting as the inverse of common image generators, image captioners are a form of feature extraction models which convert images into text, functioning at the crossroads between computer vision and natural language processing [10].

Our chosen image captioner model is the *CLIP Interrogator* [11], based on Salesforce's *BLIP* model. This baseline is a multi-task model which is capable of both image understanding and image generation, and can operate in three possible modes: an unimodal encoder, an image-grounded text encoder, and an image-grounded text decoder.

Specifically, the model's *captioner* is an image-grounded text decoder. Its intent is to generate synthetic captions T_s given training images I_w collected from web datasets. In the BLIP learning framework, this is combined with the *filter*, an image-grounded text encoder which removes texts that are predicted to not match a given image – this is applied to both synthetic captions T_s and real captions T_w found inside training datasets. This is combined with a set $\{(I_h, T_h)\}$ of human-annotated images and texts to produce a robust training dataset for a ML algorithm [12]:

$$D = \{(I_w, T_w)\} + \{(I_w, T_s)\} + \{(I_h, T_h)\} \tag{1}$$

Our method works primarily with this model's captioner. We opt to use the BLIP-Image-Captioning-Large sub-model, pre-trained on the COCO dataset to produce human-readable captions for input images. However, human-readable captions are not the most effective way to produce the inputs for an image generator. An image generator's input text, known as a "prompt", needs to be detailed and describe multiple keywords pertaining to an image: the subject of the image itself (which is typically produced by an image captioner out of the box), as well as the style, resolution, color, lightning and other details.

To that end, we use the CLIP Interrogator to first generate a baseline caption using BLIP, and then simplify the caption while adding additional keywords which most closely match the target image. These come from a predefined dataset known as "flavors", and include keywords such as "highly detailed", "sharp focus", "intricate", "digital painting" as well as phrases referring to specific objects and entities located within an image. The Interrogator selects the most appropriate keywords and phrases from "flavors" dataset by measuring the distance between a target image and each separate phrase.

3.2. Synthesizing images from text prompts

After the images within an adapted print publication are converted to text prompts, the next step is to transfer them to an AI image generator such as *Stable Diffusion* to obtain a new set of images designed with accessibility in mind. Since the CLIP Interrogator's prompt generator has been designed with Stable Diffusion in mind, its newest stable version, SDXL v 1.0, has been selected as the image generator of choice [13]. Its open-source nature means that it can be deployed on any local machine as part of our overall method.

This image generator is a *latent diffusion model*, an improvement on traditional diffusion models. Traditional diffusion models work by first "corrupting" training data, such as images, by adding noise to their inputs in a step-by-step-process. At each time step, Gaussian noise is added to a data distribution $x_0 \sim q(x_0)$ with variance $\beta_t \in (0, 1)$, resulting in the iterative process over the distribution of the variable:

$$q(x_1, \dots, x_T \mid x_o) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}) \tag{2}$$

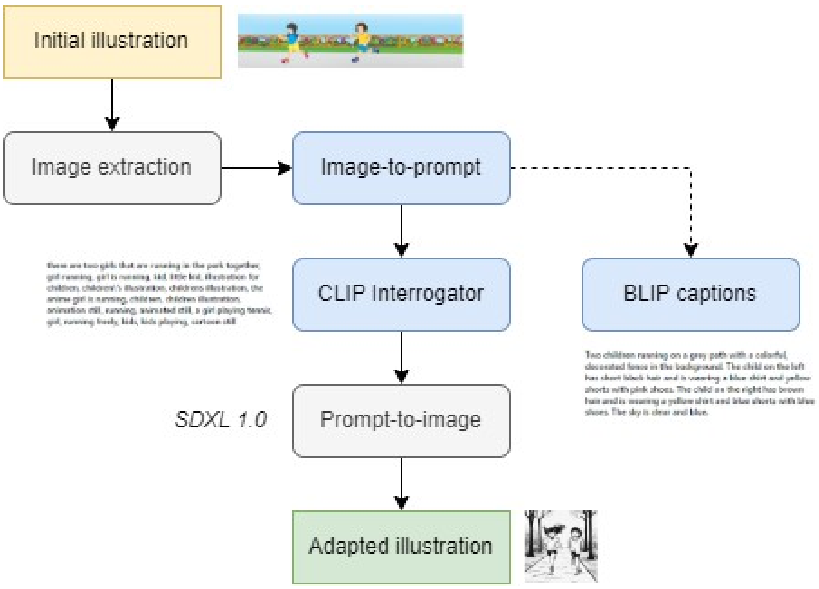
$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I) \tag{3}$$

This process is called *forward diffusion*, and concludes once the distribution q is sufficiently similar to pure Gaussian noise.

Reverse diffusion is the process of recovering the original image from the resulting noise. The overall workflow of diffusion models after they have been trained is to generate new images, given random noise as input. Latent diffusion models perform this process within *latent space* – a mathematical representation of data where similar items are grouped.

Aside from a shortened version of the CLIP Interrogator's keyword-based output, we append several keywords designed at simplifying and matching them to a more clear and simplified style, such as "illustration for children", "monochrome", "very low detail" and "no shading". This ensures that, while the objects and entities denoted by the keywords are included within the final image, it remains simplified without obstructing valuable information by noisy elements. Keywords can be modified or adjusted as needed – for instance, "monochrome" may be removed should we require a full-color illustration.

The complete workflow of our adaptation method can be seen in Figure 1.



there are two girls that are running in the park together, girl running, girl is running, kid, little kid, illustration for children, children's illustration, childrens illustration, the anime girl is running, children, children illustration, animation still, running, animated still, a girl playing tennis, girl, running freely, kids, kids playing, cartoon still

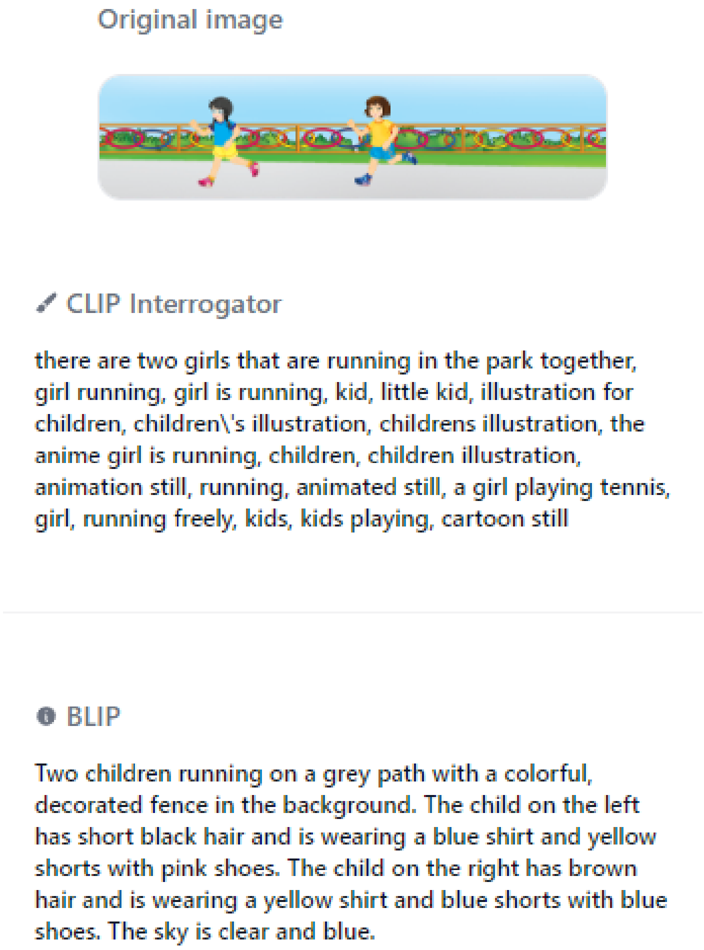
Two children running on a grey path with a colorful, decorated fence in the background. The child on the left has short black hair and is wearing a blue shirt and yellow shorts with pink shoes. The child on the right has brown hair and is wearing a yellow shirt and blue shorts with blue shoes. The sky is clear and blue.

Figure 1: The illustration adaptation method's overall workflow.

4. Results and Discussion

In Figure 1, we can see the results of this image-to-text generation model applied to one of the illustrations from a selected Ukrainian adapted textbook. Initially, the caption generated by the model is a verbose human-readable representation; the CLIP Interrogator modifies this to an image generator based on matching keywords from a preset list, which is less human-readable, but more effective as an input for a future image generation step.

The number of keywords within the text prompt can be modified at will by restricting the number of phrases returned within the Interrogator, prioritizing those which most closely match the image. While a single image-to-text conversion is shown on Figure 1, our method is a batch process. This means that after all images have been extracted from a printed publication's source file, they are converted into textual prompts with no additional human interference, significantly speeding up the captioning process as compared to a human captioner.



Original image

CLIP Interrogator

there are two girls that are running in the park together, girl running, girl is running, kid, little kid, illustration for children, children\'s illustration, childrens illustration, the anime girl is running, children, children illustration, animation still, running, animated still, a girl playing tennis, girl, running freely, kids, kids playing, cartoon still

BLIP

Two children running on a grey path with a colorful, decorated fence in the background. The child on the left has short black hair and is wearing a blue shirt and yellow shorts with pink shoes. The child on the right has brown hair and is wearing a yellow shirt and blue shorts with blue shoes. The sky is clear and blue.

Figure 2: An example of image-to-text generation with both human-readable captions (intended for "visual question answering" tasks), as well as prompts for image generators.

An example of image generations, which are the final output of our method, can be seen in Figure 3. Within this example, the leftmost image, also shown separately as Figure 4, is particularly appropriate for accessible publications, as it creates an image in a simplified cartoon style with minimal shading that still keeps its core elements (the girls in the foreground and trees in the background) legible. This can be translated directly into a mixed-format illustration which combines visual elements with Braille-like tactile dots.

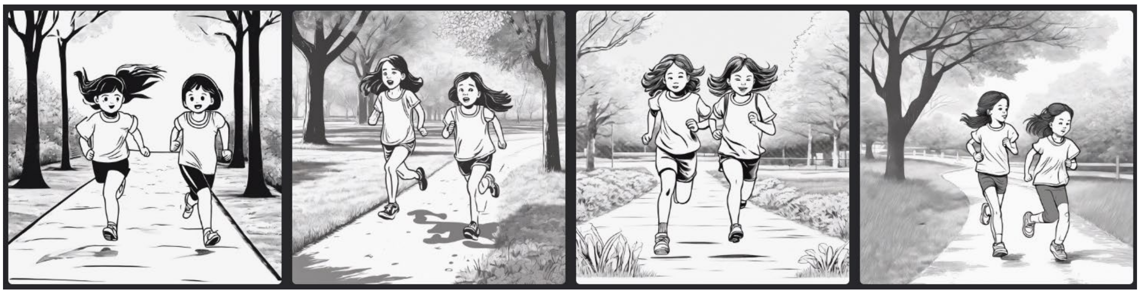


Figure 3: An example of text-to-image generation, and the model's final result. Prompt: "There are two girls that are running in the park together, girl running, girl is running, little kid, illustration for children, monochrome, very low detail, no shading, simple line art".

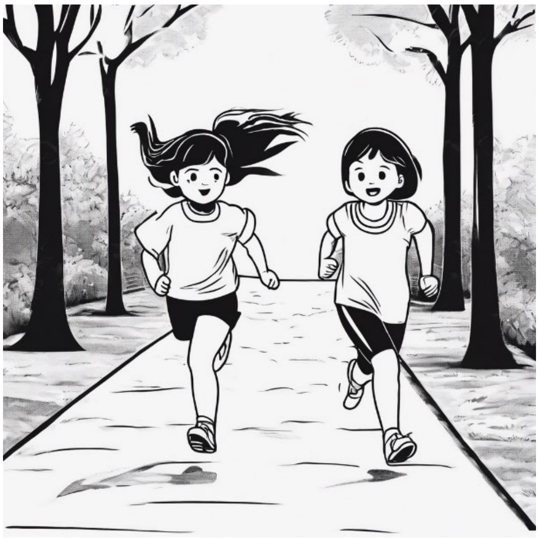


Figure 4: The generated image chosen as the final adapted illustration.

The software implementation of our method uses the Python-based Jupyter Notebook environment to integrate several steps of the process: the PyMuPDF library is used to extract images from a printed publication's source file; the CLIP Interrogator (internally based on the PyTorch library & its torchvision extension) is responsible for extracting prompts for images; while the SDXL 1.0 generator is used to create new illustrated images. On a RTX 4090 GPU, our pipeline takes ~22 seconds to convert an original illustrated image into four generated variants ready to be reviewed by a publication's editor, meaning that the entirety of a print publication's illustrations can be regenerated within hours.

5. Conclusions

This article describes a method for adapting illustrations within the source file print publication, with no or minimal human supervision, that can significantly speed up the process of making a publication more accessible to a wide audience of readers, such as people with vision impairment. This method enables both educational and volunteer organizations to produce high-quality illustrations for an adapted print publication.

Potential areas for further research include improvements to the prompts used within our image generation step – for instance, adding support for multi-colored, yet clean and simplified illustrations. The image generation model itself also has potential for improvement; as AI image generation is a rapidly developing field of research, our machine learning pipeline can be periodically revisited to make use of the newest models and techniques.

References

[1] M. Faverio, A. Perrin, Three-in-ten Americans now read e-books, Pew Research Center, 2022. URL: https://www.pewresearch.org/short-reads/2022/01/06/three-in-ten-americans-now-read-e-books/.

[2] A. Amirtharaj, D. Raghavan, J. Arulappan, Preferences for printed books versus E–books among university students in a Middle Eastern country, Heliyon (2023) e16776. doi: 10.1016/j.heliyon.2023.e16776.

[3] R. Zatserkovnyi et al., Application for Determining the Usability of Adapted Textbooks by People with Low Vision, in: Proceedings of the 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 19–21 October 2023. doi: 10.1109/csit61576.2023.10324055.

[4] L. Salmerón et al., Reading comprehension on handheld devices versus on paper: A narrative review and meta-analysis of the medium effect and its moderators, Journal of Educational Psychology (2023).

[5] Enhancing Reach: The Fundamentals of eBook Accessibility, Ingram Content Group, 2024. URL: https://www.ingramcontent.com/publishers-blog/fundamentals-of-ebook-accessibility .

[6] A. Mangen, A. Weel, The evolution of reading in the age of digitisation: an integrative framework for reading research, Literacy (2016) Vol. 50, no. 3, pp. 116–124. doi: 10.1111/lit.12086.

[7] D. Valente et al., Comprehension of a multimodal book by children with visual impairments, British Journal of Visual Impairment (2023) 42(2), 026461962311720. doi: 10.1177/02646196231172071.

[8] K. Zebehazy, A. Wilton, Graphic Reading Performance of Students with Visual Impairments and Its Implication for Instruction and Assessment, Journal of Visual Impairment & Blindness (2021) Vol. 115, no. 3, pp. 215–227. doi: 10.1177/0145482x211016918.

[9] A. Kuzmin, O. Pavlova. Analysis of Artificial Intelligence Based Systems for Automated generation of Digital Content, Computer Systems and Information Technologies (2024) no. 1, pp. 82-88. doi: 10.31891/csit-2024-1-10

[10] H. Udo, T. Koshinaka, Image Captioners Sometimes Tell More Than Images They See, arXiv.org, 2023. URL: https://arxiv.org/abs/2305.02932.

[11] pharmapsychotic/clip-interrogator: Image to prompt with BLIP and CLIP, GitHub, 2024. URL: https://github.com/pharmapsychotic/clip-interrogator.

[12] J. Li et al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, arXiv.org, 2022. URL: https://arxiv.org/abs/2201.12086.

[13] Stable Diffusion XL - SDXL 1.0 Model, 2024. URL: https://stablediffusionxl.com/.