Generation of realistic children's book images based on diffusion models

Generación de imágenes realistas de libros infantiles basadas en modelos de difusión

Nayeli Joaquinita Meléndez Acosta[1, 2*] https://orcid.org/0000-0001-9320-0834
Edmundo Bonilla Huerta[2] https://orcid.org/0000-0003-2062-4219
José Federico Ramírez Cruz[2] https://orcid.org/0000-0002-4468-4171
Yesenia Nohemí González Meneses[2] https://orcid.org/0000-0003-1034-0204

Recibido 5 de abril de 2024, aceptado 28 de mayo de 2024
Received: April 05, 2024 *Accepted: May 28, 2024*

ABSTRACT

This paper describes a model for generating realistic images for children's books using diffusion models and explains each step of the proposed model. However, the focus is on reviewing the existing area of research on controlled diffusion models. The focus is on ControlNet, a neural network model for controlling stable diffusion models that allows the addition of controls to guide image generation. In this work, edges and sketches have been used as controls. The edges have been tested using the Canny, Sobel and Prewitt edges. Furthermore, the project aims to design a bank of images extracted from children's books to generate new versions of books with realistic images. The ControlNet pre-training model has been used to create realistic images of children's books, specifically classic storybooks; our tests have been carried out with the books 'The Little Prince' and 'The Seagull and the Penguin'. The research findings bear significant practical implications, affirming that ControlNet is a potent tool for generating visually attractive, diverse, and high-quality images. This reassures illustrators and developers working on children's books about the relevance and applicability of this invaluable resource. In addition, it was concluded that using edges as a control in image generation is better, as edges provide more detail of the original image.

Keywords: Realistic images, children's books, diffusion models, ControlNet, edges.

RESUMEN

En este artículo se presenta un modelo para la generación de imágenes realistas para libros infantiles usando modelos de difusión, se explica cada etapa del modelo propuesto, pero nos enfocamos en la revisión del área de investigación existente sobre los modelos de difusión controlados. En particular nos enfocamos en ControlNet. ControlNet es un modelo de red neuronal para controlar modelos de difusión estable, lo que permite agregar controles para guiar la generación de imágenes. Se utilizan los bordes y bocetos como control, en el caso de los bordes se prueban los bordes de Canny, Sobel y Prewitt. También se pretende diseñar un banco de imágenes extraídas de libros infantiles, para generar nuevas versiones de libros con imágenes realistas. Se usa el modelo pre entrenado ControlNet para crear las imágenes realistas para libros infantiles, específicamente libros de cuentos clásicos, nuestras pruebas

[1] Universidad del Istmo. Campus Ixtepec. Oaxaca, México. E-mail: nayelim@bianni.unistmo.edu.mx
[2] Tecnológico Nacional de México. Campus Apizaco. Tlaxcala, México.
E-mail: edmundo.bh@itapizaco.tecnm.mx; federico.rc@itapizaco.tecnm.mx; yesenia.gm@itapizaco.tecnm.mx
* Autor de correspondencia: nayelimelendez@gmail.com

se realizan utilizando los libros: "El principito" y "La gaviota y el pingüino". Los resultados muestran que ControlNet genera imágenes visualmente atractivas, diversas y de muy buena calidad, por lo que es un buen generador de ilustraciones para libros infantiles, además se concluye que es mejor utilizar los bordes como control en la generación de imágenes, debido a que los bordes proporcionan más detalles de la imagen original.

Palabras clave: Imagen realista, libros infantiles, modelos de difusión, ControlNet, bordes.

INTRODUCTION

Currently, diffusion models for synthetic image generation have shown remarkable performance in generating visually appealing realistic images [1], [2]. Thus, text-to-image models have shown unprecedented capabilities by generating images based on a brief descriptive message entered by the user [3], [4]. However, many times, this control based on messages alone is not enough to meet the needs because frequently, the text descriptions do not provide information to control the results of the final generation [2], [5], [6].

For more customized image generation, there is a need to introduce or incorporate additional controls to the text-to-image models, such as edges maps, sketch drawn by the user, segmentation maps, depth maps, and others., beyond the text that allow controlling the output [7]-[9]. The build quality and additional controls make the text-to-image diffusion models suitable for real-world applications such as fashion design, interior decoration, early childhood education. They can also be easily applied to tasks such as painting and coloring [6], [7], [10].

Children's picture books are a tool for vicarious learning and provide important content that increases cultural awareness, language awareness, and motivation [11]. Thus, generating images from text can help humans create diverse visual content very easily. Motivated by the competition of diffusion models to generate realistic images, these models were applied to generating images for children's books.

The creation of images as illustrations for children's books is very laborious since it requires specialized skills, a well-trained artist or illustrator, and hours of work [12]-[14]. Therefore, creating a bank of images that serve to illustrate the texts of the book using a generator tool can help create diverse visual content with great ease [7].

The main objective is to describe a model for automatically generating realistic images for children's books. After reviewing existing text-to-image diffusion models, is performed, a propose the use of a ControlNet conditioned text-to-image diffusion model to generate realistic images using Canny, Sobel, and Prewitt edges as control.

LITERATURE REVIEW

Text-to-image controlled diffusion models have been very successful in generating images. These models require a brief descriptive message entered by the user. Additionally, they are provided with an image to control the output to generate a controlled image that is visually attractive. Some works related to these models and that are relevant to our research are:

- Zhang *et al.* [2] present ControlNet, a neural network structure that allows controlling a pre-trained diffusion model through specific input conditions. ControlNet allows the generation of an image based on specific conditions; some conditionals they consider in their tests are Canny edges, segmentation, user scribbles, depths, and Hough lines. They perform their tests with both small and large data sets. Conditionals can enrich and facilitate the use of diffusion models in more applications.
- Aggarwal *et al.* [5] explore image generation using a two-step process: generating a CLIP image from text (a Diffusion Prior) and generating an image from the CLIP image (a decoder). By training Diffusion Prior with additional conditional specific information, such as the color histogram, one can control the generation of images in to a desired domain.
- Nichol *et al.* [7] propose GLIDE, a guided language system for generation and editing using image diffusion; they compare two different guiding strategies: CLIP and classifier-free guiding. In their tests, human judges evaluate

them. The guide without a classifier is preferred both for the realism of the images and for its similarity with the subtitles. Furthermore, they found that this strategy produces higher-quality images and reflects a wide range of world knowledge.

- Rombach *et al.* [10] propose to add layers of cross-attention in latent space in the architecture of a pre-trained diffusion model. These models consume hundreds of GPU days since they operate directly in pixel space, and the inference has a high computational cost due to sequential evaluations. Its latent diffusion model (LDM) achieves state-of-the-art scores in adding color to images and conditional image synthesis; it also reduces computational requirements compared to pixel-based diffusion models (DM).
- Zhao *et al.* [6] present an approach that allows Uni-ControlNet to simultaneously use different controls in pretrained text-to-image (T2I) diffusion models. Uni-ControlNet distinguishes two groups of conditions: local conditions (such as depth maps, user sketches, edges maps, and segmentation) and global controls (such as CLIP image embeds).Only two adapters are added regardless of the number of conditions to handle both global and local conditions, significantly reducing the training cost. The results show that Uni-ControlNet outperforms existing methods regarding control and generation quality.
- Huang *et al.* [9] present Composer as a new paradigm for image generation that allows controlling the output image. The composer is a multi-conditional diffusion model composed of generative models that can seamlessly recombine visual components to produce new images. Composer first decomposes an image into representative factors; then, factors are used as conditions to train the diffusion model. Composer, like Uni-ControlNet, accepts global conditions like text descriptions and local conditions like depth maps or sketches entered by the user.
- Mou *et al.* [8] propose Adapter-T2I a generative model that learns adapters. this allows to order the internal knowledge of the T2I model through external control conditions. Thus, it is possible to train several adapters according to different conditions to achieve the desired control. T2I-Adapter provides previously established T2I diffusion models with additional guidance. Their experiments show that Adapter-T2I generates quality images and widely applicable.

It is important to decide which model might be the most suitable for generating images for children's books. Table 1 compares ControlNet, a controlled fusion generator model, with some others in terms of the method used, the size of the database, the capabilities of the model, and its accessibility.

METHODOLOGY

Producing a book with rich and diverse visual content is essential to generate a wide variety of images to embellish the book and enhance its content. So, it is important to have a tool capable of generating images that can be controlled in detail over the output.

The block diagram of the proposed model (Figure 1) for generating children's books images using an edge-controlled diffusion model of the images is shown here. The first stage involves collecting books in PDF. The next stage involves extracting metadata from the collected books, and the third stage involves collecting images. In the next stage, edge detection is performed, and the resulting image acts as a controller for the diffusion model. In the fifth stage, images are generated using the

Table 1. Comparative summary of some generative models.

Model	Method	Database size	Open source	Capacities
DALL-E 2	Zero-Shot	650M	Partially	Text to Image
GAN	Unsupervised	Several	Partially	Image to Image
Mini DALL-E	Transformer	3M, 12M y 15M	Yes	Text to Image
ControlNet	Neural network blocks	Several	Partially	Text/Image to Image

Source: Adapted from [15].

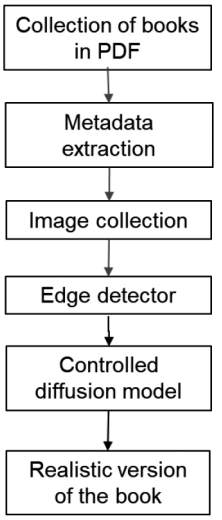

Figure 1. Block diagram of the proposed model for the generation of realistic images for children's books using a controlled diffusion model.

ControlNet diffusion model. Finally, in the last stage, the images in the book are updated.

Collection of books in pdf

In this stage, the books used in the tests are collected, and the illustrations are compiled manually by searching the Internet for books in PDF format. For now, we focus on illustrated books of classic tales such as the Little Mermaid, Little Red Hood, Alice in Wonderland, the Little Prince, Around the World, and The Wizard of Oz, to name a few.

Metadata extraction

The construction of the image metadata table is performed at this stage. Metadata extraction is performed using pdfimages, this tool provides a list of all the images contained in a PDF file, along with various metadata. Table 2 shows the metadata table of the book "The Little Prince," which is composed of eight metadata: 1) image number, 2) image name, 3) image location page, 4) the two dimensions (width, height) of the image, 5) image format, 6) object number (location of the image in the pdf) and 7) the image description tag.

Image collection

In this stage, the PDF books collected in the first stage are used to collect images. In this stage, the pdfimages tool is also used because it allows us to extract the files of all the images in their original sizes and formats, creating the database of images with them.

Edge detector

There are different conditions that allow diffusion models to control output such as canny edges, user sketches, normal shapes, depths, and others. [2]. In this work we have used three types of edges: Canny, Sobel, and Prewitt to control the diffusion model. Edge detection was used because they preserve the structural properties of an image.

Canny operator. The first edge detector tested against to compare our edge detector is Canny, as this detector is known as the optimal edge detector [16].

Sobel operator. This operator is widely used as an edge detection method. Sobel is often used as a

Table 2. Metadata of the images of the book "The Little Prince."

#	name	page	width	Height	format	object	label
1	image001	7	257	158	Jpg	26	boa eating a mouse realistic image.
2	image002	7	299	98	Jpg	27	boa with an elephant in the stomach realistic image.
3	image003	11	338	365	Jpg	38	the little prince realistic image.
4	image004	12	129	114	Jpg	41	white lamb jumping realistic image.
5	image005	12	135	127	Jpg	42	white lamb realistic image.
6	image006	12	135	116	Jpg	43	white lamb realistic image.
7	image007	12	146	62	Jpg	44	white box with three holes realistic.
8	image008	15	262	357	Jpg	51	the little prince on the asteroid B 612 realistic image.
40	image040	82	278	272	Jpg	222	the little prince sitting on the sand realistic image.
41	image041	84	476	660	Jpg	228	the little prince gently falling into the sand realistic image.

Source: Adapted from [15].

simple horizontal and vertical edge detector. The operator consists of a pair of 3 × 3 convolution kernels, one of them is rotated 90º [16], [17].

Prewitt Operator. The Prewitt operator, like the Sobel operator, detects vertical and horizontal edges in images. The gradient is estimated in eight possible directions (for a 3 × 3 convolution mask), with the largest magnitude convolution result indicating the direction of the gradient [16].

Controlled diffusion model

In many real-world applications, you cannot rely solely on text prompts as flexible and precise control is required. Therefore, the ControlNet diffusion model has been used to generate realistic synthetic images for our books' new versions. ControlNet not only allows for the addition of control but also ensures the output is adjusted to a pre-trained text-to-image model with task-specific conditions. In our case, this control is crucial for the generation of images.

ControlNet is a network architecture that can enhance previously trained image diffusion models. It manipulates the input conditions of neural network blocks to further control the overall behavior of the entire network. The control used in this work are the edges, and the operators used in the tests were already explained in the previous stage.

Realistic version of the book

In this last stage, a new version of the book is generated in the following way, the illustrations of the original book are updated with the new realistic images generated using the ControlNet text-to-image diffusion models using three edge detectors as controls: Canny, Sobel, and Prewitt. To carry out the update, it is necessary to consult the metadata table to obtain the object number, which is an identifier that allows us to know the location of the image within the pdf.

RESULTS

Synthetic images are generated using the image edges as conditionals with the ControlNet image model. This section shows our experiments generating realistic images for a new book version; the implementation is done in Python. Two books: "The Little Prince" and "The Seagull and the Penguin" have been used for the tests. It is important to mention that once the new image is generated, it must be resized to match the original image's size. Here are some results of rendering realistic images using edge maps or user sketches.

Figure 2 and Figure 3 show the generation of the images using the ControlNet model. For the test, images 3, 10, 30, and 38 of the book "The Little Prince", and images 2, 4, 7, and 8 of the "The Seagull and the Penguin" book have been selected. In these tests, the Canny, Sobel, and Prewitt edges as conditionals have been used; the text prompt, which is the label of the image in the metadata table is also provided. The results show that the visual quality of the generated images is very good. Furthermore, the edges are much better defined using the Canny operator, which is known as the optimal edge detector.

On the one hand, it was possible to detect that sometimes the elements that make up the illustration are not generated correctly, mainly in the hands and face, such as the eyes, mouth, and nose, but sometimes the facial expressions are not natural.

On the other hand, the generation of any other object is very good, which is favorable since it is essential to have a great variety of illustrations for books. This proves that synthetic images are also a good way to improve diversity in books, and the creation of books.

Figure 4 shows the results of using a sketch created by the user to control the output using the ControlNet broadcast model. The labels selected for the images to be created are image 3 from the book "The Little Prince" and image 2 from the book "The Seagull and the Penguin."

From Figure 4's output, it can be seen that ControlNet generates realistic images very similar to the sketch provided as input. In comparing both controls used in this work -edges and sketches- the edges are more precise than the sketches, but their usefulness will depend on the problem to which they are applied. Our research on the generation of realistic images for children's books has yielded a practical recommendation: it is better to use edges as a control in image generation. This approach, by providing more details from the original image, can significantly enhance the quality and realism of the illustrations, making them more engaging for young readers.

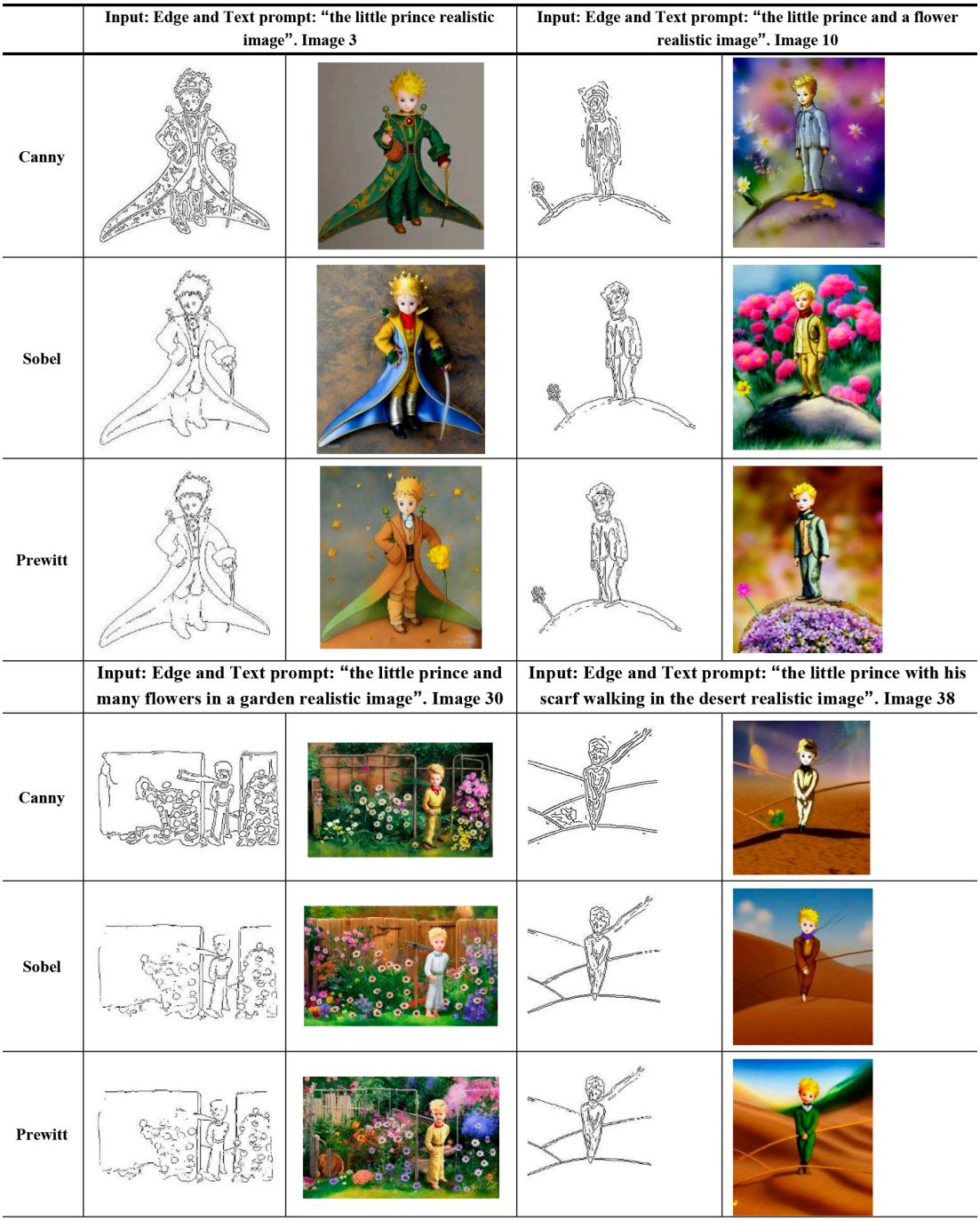

Figure 2. “The Little Prince” book. Results of using the diffusion model: ControlNet. We used to control the output: the edges obtained by using the cellular automaton, Sobel edges, Roberts edges and Prewitt edges.

On the one hand, it was considered that all models allow enhancing the image bank for future work, but if a more controlled generation of images in terms of content using the edges is desired, ControlNet should be used. On the other hand, ControlNet also allows controlling the coherence using an image as

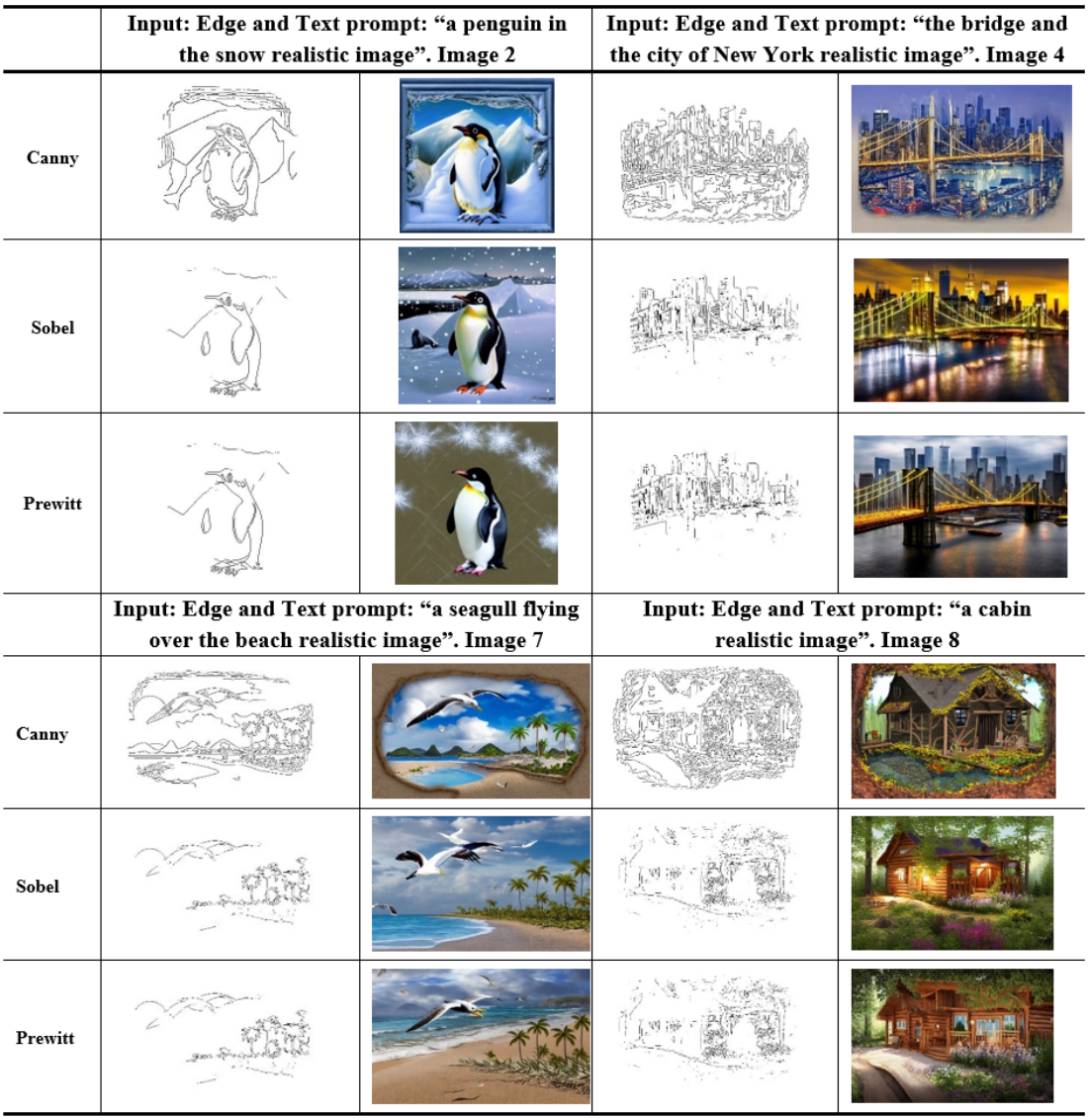

Figure 3. Book "The seagull and the penguin." Results of using the diffusion model: ControlNet. We used to control the output: the edges obtained by using the cellular automaton, Sobel edges, Roberts edges and Prewitt edges.

a reference and as detailed input text as possible. For now, this work has only addressed control in terms of content. It is worth mentioning that the images are generated randomly, but the control is established from the metadata because we know what realistic or hyperrealistic image will be generated. Everything has been programmed in Python due to the accessibility of the diffusion generator models, but it is important to mention that DALL-E 2 and ControlNet only allow the generation of a limited number of images.

Figure 5 shows the updating of the images on pages 11 and 28 of the book "The Little Prince." Figure 5a, shows page 11 of the original book, Figure 5b shows page 11 updated with a new realistic image generated with the ControlNet model using Canny's edges as output control. In Figure 5c, you can see page 28 of the original book. On the left, in Figure 5d, page 28 is shown updated with a new realistic image generated with the ControlNet model using Canny's edges as control.

Input: sketch and Text prompt: "the little prince realistic image". Image 3. Book: "The little Prince"		Input: Sketch and Text prompt: "a penguin in the snow realistic image". Image 2. Book: "The seagull and the penguin"	
			

Figure 4. Results of using the diffusion model: ControlNet. We use to control the output: user sketch.

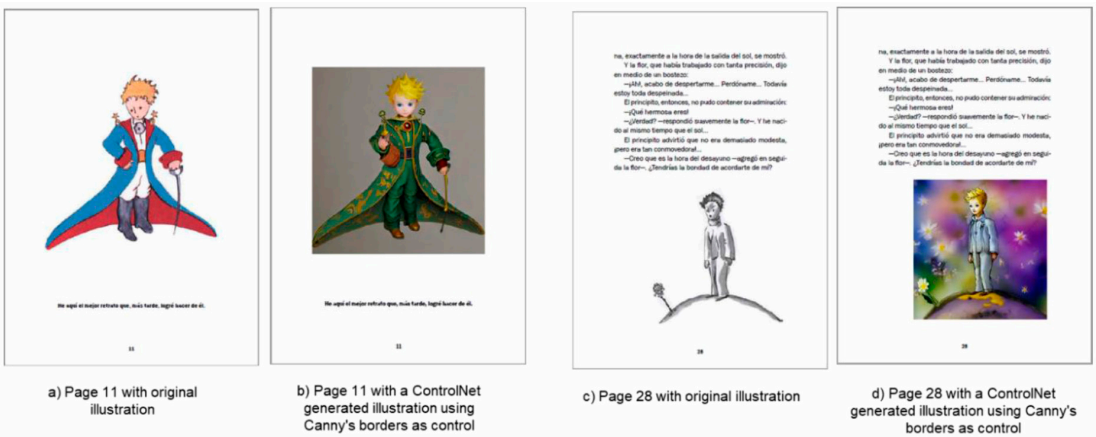

Figure 5. Update of page 11 of the book "The Little Prince." a) original page, b) new page with a realistic image generated using the ControlNet model controlled by the Canny operator. Update of page 28 of the book "The Little Prince." c) original page, d) new page with a realistic image generated using the ControlNet model controlled by the Canny operator.

Finally, Figure 6 shows the updates on pages 3 and 11 of the book "The Seagull and the Penguin." Here, image variability is a significant element in books, as an image can change its presentation. Figure 6a shows page 3 of the original book and Figure 6b shows page 3 updated with a realistic image generated with the ControlNet model using the Sobel edge map. In Figure 6c you can see page 28 of the original book and on the left in Figure 6d is shown updated page 28 with a new realistic image generated with the ControlNet model using Canny edges.

CONCLUSIONS

The automatic generation of illustrations is a very difficult task, but it can be done using natural language or a reference image. Following a review of some existing models for automatic image generation, a model that is adaptable and most suitable for generating illustrations for children's books has been selected.

The experimental results indicate that ControlNet is a good option to be used as an illustration generator

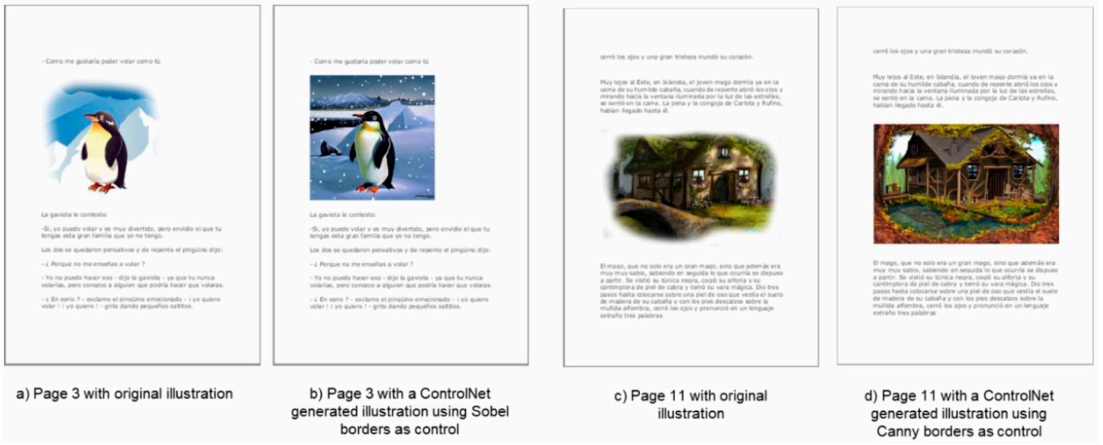

Figure 6. Update of page 3 of the book "The seagull and the penguin." a) original page, b) new page with a realistic image generated using the ControlNet model controlled by the Sobel operator. Update of page 11 of the book "The seagull and the penguin." c) original page, d) new page with a realistic image generated using the ControlNet model controlled by the Canny operator.

tool to use a short text message and an image as a reference. In the carried out tests used two controls as a reference image: edges maps and sketches have been used. Three operators have been used to generate the edge image: Canny, Sobel and Prewitt. The results show that the three operators provided promising results in image generation, although it should be noted that Canny identifies more edges than Sobel and Prewitt, which allows to have even more control over the image generation. The sketches generate images like the original, but the resulting image needs further details. Both controls allow increasing the image bank for future work, but it is better to use the edges as a control if generated images containing more detail than the original image are required.

Furthermore, future work is intended to utilize a proprietary edge detector to control the image content and also test the use of other controls, such as key points of human poses, with the goal of identifying the best control to obtain the image content. Additionally, the generated images are intended to be used to build a database of synthetic images, which can be used to create a book with dynamic images. In this way, the user can choose which illustration he/she wants to see from a set through augmented reality.

This tool will also allow for more significant interaction with the children, allowing them to make their sketches and use these tools to generate the illustrations. It is also pertinent to mention the importance of investigating the reading motivation of children and carrying out samplings in schools with the support of pedagogues and psychologists to see the reactions of children when reading and seeing a book with dynamic illustrations.

REFERENCES

[1] A. Stöckl, "Evaluating a synthetic image dataset generated with stable diffusion," i*n Proceedings of Eighth International Congress on Information and Communication Technology, ICICT 2023, Lecture Notes in Networks and Systems*, X.S. Yang, R.S. Sherratt, N. Dey, A. Joshi, Eds. vol. 693, 2023, pp. 805-818, doi: 10.1007/978-981-99-3243-6_64.

[2] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," i*n 2023 IEEE/CVF International Conference on Computer Vision (ICCV),* Paris, France,2023, pp. 3813-3824, doi: 10.1109/ICCV51070.2023.00355.

[3] N. Ruiz, Y. Li, V. Jampan, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," i*n 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Vancouver,

BC, Canada, 2023, pp. 22500-22510, doi: 10.1109/CVPR52729.2023.02155.

[4] G. Rinon, Y. Alaluf, Y. Atzmon, O. Patashnik, A.H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: personalizing text-to-image generation using textual inversion," 2022, arXiv:2208.01618.

[5] P. Aggarwal *et al.*, "Controlled and conditional text to image generation with diffusion prior," 2023, arXiv: 2302.11710.

[6] S. Zhao *et al.*, "Uni-ControlNet: All-in-one control to text-to-image diffusion models," *37th Conference on Neural Information Processing Systems*, 2023, pp. 11127-11150.

[7] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," *In International Conference on Machine Learning*, vol. 162, Baltimore, Maryland, USA, 2022, pp. 16784-16804. [Online]. Available: https://proceedings.mlr.press/v162/nichol22a/nichol22a.pdf

[8] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2I-Adapter: Learning adapters to Dig out more controllable ability for text-to-image diffusion models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, pp. 4296-4304, 2024, doi: 10.1609/aaai.v38i5.28226.

[9] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," *Proceedings of the 40th International Conference on Machine Learning, 2023,* pp. 13753-13773. [Online]. Available: https://proceedings.mlr.press/v202/huang23b/huang23b.pdf

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution image synthesis with latent diffusion models," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 10674-10685, doi: 10.1109/CVPR52688.2022.01042.

[11] A. Fatimah, S. Santiana, and Y. Saputra, "Digital Comic: an innovation of using toondoo as media technology for teaching ingles short story," *English Review: Journal of English Education*, vol. 7, no. 2, 2019, doi: 10.25134/erjee.v7i2.1526.

[12] J. Chen, G. Liu, and X. Chen, "AnimeGAN: A novel lightweight GAN for photo animation," *in Artificial Intelligence Algorithms and Applications, ISICA 2019. Communications in Computer and Information Science*, K. Li, W. Li, H. Wang, and Y. Liu, Eds. vol. 1205, 2020, doi: 10.1007/978-981-15-5577-0_18.

[13] Y. Shu *et al.*, "GAN-Based multi-style photo cartoonization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 10, pp. 3376-3390, 2022, doi: 10.1109/TVCG.2021.3067201.

[14] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised Dual Learning for Image-to-Image Translation," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2868-2876, doi: 10.1109/ICCV.2017.310.

[15] N.J. Meléndez, E. Bonilla, J.F. Ramírez y Y.N. González, "Generador de ilustraciones para libros utilizando inteligencia artificial," Research *in Computing Science*, vol. 152, no. 9, pp. 51-65, 2023. [Online]. Available: https://rcs.cic.ipn.mx/2023_152_9/Generador%20de%20ilustraciones%20para%20libros%20utilizando%20inteligencia%20artificial.pdf

[16] M. Sonka, V. Hlavac, and R. Boyle, "Image pre-processing." in *Image Processing: Analysis and Machine Vision*, 4th ed., Stamford, CT, USA: Cengage Learning, 2015, pp. 116-174.

[17] M. Raman and A. Himanshu, "Study and comparison of various image edge detection techniques," *International Journal of Image Processing (IJIP),* vol. 3, no. 1, pp. 1-11, 2009.