Illustrating Classic Brazilian Books using a Text-To-Image Diffusion Model

Felipe Rodrigues Perche Mahlow, André Felipe Zanella, William Alberto Cruz Castañeda, and Regilene Aparecida Sarzi-Ribeiro

***Abstract*—In recent years, Generative Artificial Intelligence (GenAI) has undergone a profound transformation in addressing intricate tasks involving diverse modalities such as textual, auditory and visual generation. Within this spectrum, text-to-image (TTI) models have emerged as a formidable approach to generating varied and aesthetically appealing compositions, spanning applications from artistic creation to realistic facial synthesis, and demonstrating significant advancements in computer vision, image processing, and multimodal tasks. The advent of Latent Diffusion Models (LDMs) signifies a paradigm shift in the domain of AI capabilities. This article delves into the feasibility of employing the Stable Diffusion LDM to illustrate literary works. For this exploration, seven classic Brazilian books have been selected as case studies. The objective is to ascertain the practicality of this endeavor and to evaluate the potential of Stable Diffusion in producing illustrations that augment and enrich the reader's experience. We will outline the beneficial aspects, such as the capacity to generate distinctive and contextually pertinent images, as well as the drawbacks, including any shortcomings in faithfully capturing the essence of intricate literary depictions. Through this study, we aim to provide a comprehensive assessment of the viability and efficacy of utilizing AI-generated illustrations in literary contexts, elucidating both the prospects and challenges encountered in this pioneering application of technology.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9172

***Index Terms*—Image Generation, Diffusion Models, Illustration, Text-to-Image**

I. Introduction

Generative Artificial Intelligence (GenAI) has revolutionized various tasks by integrating capabilities in text, audio, video, and image generation. GenAI excels in creating synthetic data that closely mimics real-world phenomena. For instance, text generation models such as OpenAI's GPT [1] have transformed the field of writing by demonstrating an exceptional understanding of context and coherence [2]. These models enhance natural language processing, content creation, and automated writing tasks [3]. In the realm of audio, models like Tacotron [4] and WaveNet [5] leverage deep neural networks to generate realistic speech and music, pushing the boundaries of audio synthesis [6]. Similarly, image generation has seen significant advancements with models such as DALL-E [7], [8], MidJourney [9], and Stable Diffusion [10], which can create intricate images from textual descriptions. Additionally, Generative Adversarial Networks (GANs) [11] have been pivotal in producing high-quality images for artistic endeavors and realistic face generation, significantly impacting computer vision and multi-modal tasks [12]–[15]. The ability of generative models to create human-like content has opened up new avenues for creative, automated, and innovative applications. However, these advancements also bring about concerns and challenges that need to be addressed in the future [16].

Text-to-image (TTI) models focus on developing methods and algorithms to create visual images from written text. A significant breakthrough in this field has been the rise of Latent Diffusion Models (LDMs) [10], which build upon the foundational principles of Diffusion Probabilistic Models (DPMs) [17]. Diffusion models apply a series of random transformations progressively to an image's probability distribution [18]. This iterative process generates detailed and realistic images while providing greater control over the creation process.

The integration of GenAI into the creative process has shown a profound and multifaceted impact. For instance, research on the use of TTI models in craft education indicates that artificial intelligence (AI) can aid ideation and visualization but also raised concerns about skill gaps, lack of materiality, and ethical implications, including biases and the impact on creativity and copyright [19]. Similarly, GenAI can increase creativity in writing tasks, showing improved quality of text outputs with AI support, despite a greater uniformity in generated creations [20]. Additionally, the application of AI in independent publishing, reveals both the AI's capability to enhance ideation and production, and the need for critical approaches to its role in preserving human craftsmanship [21]. AI use also has environmental implications, with research showing that text and illustration production with AI can generate significantly fewer carbon emissions compared to human methods [22]. However, the adoption of AI in creative fields is not without challenges, including ethical issues and the need for a careful balance between automation and human authorship. Analyzing these dynamics is crucial to understanding how AI can be a powerful tool in enhancing the creative process while addressing the challenges associated with its integration.

The associate editor coordinating the review of this manuscript and approving it for publication was Carlos Thomaz (*Corresponding author: Felipe Rodrigues Perche Mahlow*).

F. R. P. Mahlow and R. A. Sarzi-Ribeiro are with São Paulo State University, Bauru Campus, Brazil (e-mails: f.mahlow@unesp.br, and regilene.sarzi@unesp.br).

A. F. Zanella is with Maringá State University, Maringá Campus, Brazil (e-mail: aft.zanella@gmail.com).

W. A. C. Castañeda is with Technologycal Federal University of Paraná, Guarapuava Campus, Brazil (e-mail: williamalberto.cruz@gmail.com).

The extant literature offers a scant examination of the intersection between AI and book illustration. Traditionally, the illustration of literary works necessitated the intervention of human artists, a process often characterized by its time-consuming nature and subjectivity. However, recent advancements in AI present opportunities for automation, potentially enhancing this process by generating illustrations that encapsulate the essence and historical context of literary works with significantly reduced resources and time compared to traditional methodologies. An example of this is the GenAI methodology based on TTI for producing assisted art that allows dataset assembling, model training and fine-tuning, and content enhancement and post-processing within a historical or cultural setting [23]. Nevertheless, the endeavor of producing visually compelling and faithful images based on literary descriptions remains inherently complex. The quality and fidelity of these generated images are critically contingent upon the specificity of prompts and the efficacy of the models employed. Furthermore, challenges persist regarding data bias and the inherent limitations of AI models themselves. The training of these models is particularly contentious, as it often involves using copyrighted images without the explicit consent of the original artists [24].

Our research explores the application of LDMs for book illustration, utilizing the Stable Diffusion (SD) model to generate visual representations based on textual prompts derived from seven distinct works of classic Brazilian literature. We employed a two-phase methodology that begins with initial image generation using Stable Diffusion XL (SDXL) Base 1.0, followed by a refinement process with SDXL Refiner 1.0. This iterative approach was designed to enhance the quality and fidelity of the generated illustrations, aligning them more closely with the literary descriptions. The aim of this work was not to achieve "perfect" illustrations but to demonstrate the feasibility of such an approach, providing a methodological framework and documenting the positive and negative aspects encountered, as well as the ethical biases inherent in GenAI.

The main contributions of the paper are summarized as follows: **1.** Introduces the feasibility of employing LDMs as a novel approach for book illustration; **2.** Highlights the importance of precise prompt formulation to generate visually appealing and contextually relevant images; **3.** Provides a detailed analysis using both quantitative (CLIP and Inception Scores) and qualitative assessments; **4.** Delineates salient constraints and inherent biases of the model; **5.** Proposes prospective avenues for further investigation at the intersection of book illustration and GenAI.

The paper is structured as follows: Section II outlines the methodology employed, starting with II-A, where we present the books, followed by II-B and II-C, where we discuss the architecture of the SDXL Model for image generation and the hardware configurations used for training and inference respectively. In II-D, we provide a detailed analysis of the image generation process, while in II-E we address the techniques applied for refining these images. Further, in II-F, we detail the large-scale generation process and the subsequent selection of the most relevant images. Section II-G is dedicated to the quantitative evaluation of the results, focusing on the CLIP Score and the Inception Score, discussed in subsections II-G1 and II-G2, respectively. Section III presents an extensive exploration of the experimental results obtained, with a qualitative approach (III-A) illustrating the effectiveness and adaptability of our approach in generating diverse and visually appealing compositions in various contexts, as well as a quantitative approach (III-B) assessing the performance of the generated images based on established metrics. Finally, Section IV synthesizes the main findings, highlights the effectiveness of the image synthesis pipeline, and discusses potential directions for future research and improvements to the proposed method.

II. Methodology

This section outlines the systematic procedure followed for generating and refining illustrations for seven classical Brazilian books using the SDXL model. Fig. 1 provides a flowchart of the methodology, covering prompt creation, image generation, refinement, and evaluations. Each stage is detailed in the subsections, highlighting their roles in creating accurate and engaging illustrations.

A. The Books

The selection of these texts was guided by the goal of respecting copyright constraints while providing a rich source for creative exercises. Therefore, only texts from Brazilian literature available in the public domain were chosen.

The selected books for this study are:

- *Senhora* (1875) by José de Alencar
- *O Cortiço* (1890) by Aluísio Azevedo
- *A Viúva Simões* (1897) by Júlia Lopes de Almeida
- *Dom Casmurro* (1899) by Machado de Assis
- *Horto* (1900) by Auta de Souza
- *Os Sertões* (1902) by Euclides da Cunha
- *O Triste Fim de Policarpo Quaresma* (1915) by Lima Barreto

These works were chosen for their literary significance and their rich descriptive passages, which provide ample material for visualization through GenAI. Each text offers unique narratives and vivid descriptions that facilitate the generation of diverse and engaging illustrations. *Senhora* and *O Cortiço*, for example, are notable for their detailed portrayal of 19th-century Brazilian society, while *Dom Casmurro* and *Os Sertões* are renowned for their deep psychological and socio-political insights. *A Viúva Simões* and *Horto* contribute with their unique thematic and stylistic elements, and *O Triste Fim de Policarpo Quaresma* is recognized for its satirical and cultural critique. The choice of these works not only ensures a diverse range of illustrative challenges but also honors the literary heritage of Brazilian authors through modern technological means.

B. Stable Diffusion

SDXL is an LDM that utilizes a sequence of denoising autoencoders $\epsilon_\theta(x_t, t)$, wherein x_t represents a noisy version of the input x, and the goal is to predict a denoised version of

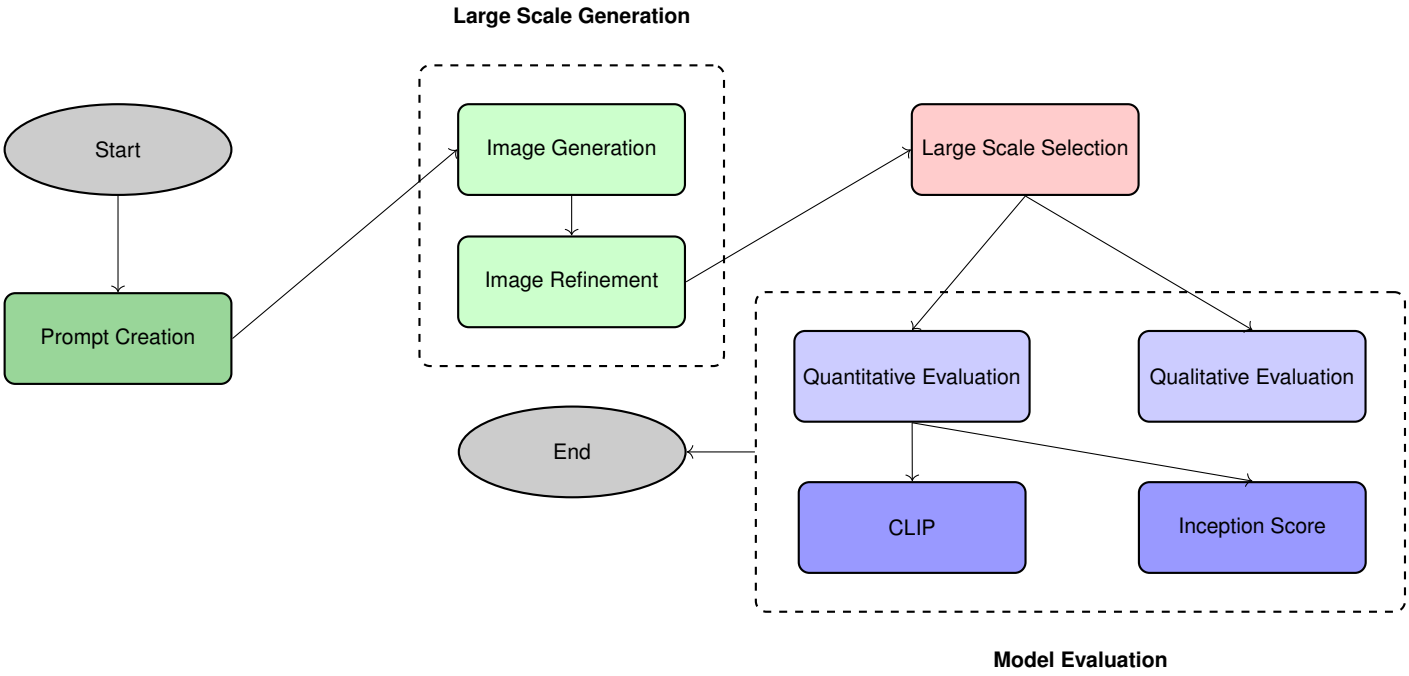


Fig. 1. Flowchart illustrating the comprehensive sequence of steps undertaken in the methodology.

the input. Consequently, generative modeling of latent representations serves as a training compression model comprising the encoder $\mathcal{E}$ and the decoder D with access to a low-dimensional latent space. This process includes the capability to construct a time-conditional UNet as a neural backbone $\epsilon_\theta(\circ, t)$, given that the forward process is predetermined, z_t obtained from $\mathcal{E}$ during training, and samples from $p(z)$ are decoded into image space through a single pass via D as described below [10], [25]:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim N(0,1),t}\left[\|\epsilon - \epsilon\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \quad (1)$$

To condition image generation on text, SDXL employs a cross-attention mechanism where the text is projected into an intermediate representation $\tau_\theta(y)$, allowing for text conditioning in the denoising process. SDXL leverages (1) a larger UNet backbone with dual text encoders (OpenCLIP ViT-bigG [26] and CLIP ViT-L [27]) for text conditioning, (2) additional unsupervised conditioning techniques, and (3) a refinement model that improves visual quality through a noising-denoising process on the latent space. Its Variational AutoEncoders (VAE) share the original Stable Diffusion design but with improved training. It was trained from scratch with a batch size of 256 (compared to 9), enhancing generalization and finer detail capture. An exponential moving average tracked weights during training for stable convergence and better performance.

C. Hardware Configuration

All computations were performed on a system equipped with an NVIDIA GeForce RTX 3090 GPU. This setup provided enough computational power and memory capacity, essential for handling the process involved in generating and refining the images.

D. Image Generation

In the first stage, the *SDXL Base 1.0*[1] [28] model was employed to generate the initial images. The model was fed with a comprehensive list of descriptive prompts, each crafted to depict specific scenes from the selected books. These prompts included detailed descriptions of characters, actions, and settings. For each prompt, the model performed the image generation in 40 inference steps. The inference process involved a denoising technique, crucial in diffusion models, which iteratively refined an image from an initial noisy state to a clear and detailed representation. In this case, the inference was configured to terminate at 0.8 of the denoising process, meaning that the image was generated to a certain level of detail before transitioning to the next stage.

E. Image Refinement

In the second stage, the *SDXL Refiner 1.0*[2] model was used to enhance the images generated in the previous step. This refinement model received the latent images (partially processed images) and continued the denoising process from 0.8, completing the remaining steps to 1.0. This additional 40-step inference phase ensured that the final images reached a higher level of quality and detail. The refiner model utilized additional components, such as a second text encoder and the VAE from the base model, to improve the accuracy and aesthetics of the images. The refiner corrected imperfections and added fine details, making the illustrations more vivid and true to the original descriptions.

[1]https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
[2]https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0

F. Large-Scale Generation and Selection

For each book, five prompts were generated, and for each prompt, 300 images, resulting in 1500 images per book. This large-scale generation ensured a wide variety of illustrations, allowing for a rigorous selection of the best representations for the project. These large quantities are important so that the calculation of quantitative metrics can be done more accurately.

G. Quantitative Evaluation Metrics

To evaluate the quality of the images generated we employed two evaluation metrics, the CLIP [29] and IS scores [30].

1) CLIP Score: metric that evaluates the semantic quality and relevance of the generated images concerning the input prompts. We utilize the *"clipq_score"* function from the *"torchmetrics.functional.multimodal"*[3] library, which calculates the CLIP score of an image concerning a text prompt. The CLIP model was trained on a large number of text-image pairs to learn a joint representation of text and image. It is capable of assessing the semantic quality of the generated images concerning the provided prompts.

2) Inception Score (IS): metric that evaluates both the quality and diversity of the generated images. A higher IS indicates that the generated images are both diverse (capturing multiple classes) and of high quality (clearly recognizable by the classifier). The metrics were calculated through implementations based on the seminal paper[45] [30].

By employing these metrics, we ensure a comprehensive evaluation of the generated images, considering both their semantic relevance to the prompts and their overall quality and diversity.

III. Results

In this section, we present and analyze examples of the generated images, focusing on specific examples that highlight the process and outcomes of our methodology. Figs. 2 and 3 showcase selected images, which facilitate a discussion on the creation process and the critical role of the prompt in generating visually captivating and contextually relevant illustrations.

While the AI system autonomously manages the image generation process, the formulation of the prompt remains a critical task for the user. Mere transcription of text from the book typically results in suboptimal outcomes. It is imperative to conceptualize the scene based on the textual descriptions provided in the book and devise a prompt that facilitates the creation of an image that is both visually compelling and faithful to the scene's intrinsic characteristics. Our analysis found patterns and anomalies in generated images, showing biases and limitations in generative models. Some images had unspecified or inconsistent elements, indicating interpretative issues. This underscores the need for diverse training data and refined models to improve contextual accuracy. Solutions may include augmenting datasets with varied examples or enhancing prompts.

[3]https://lightning.ai/docs/torchmetrics/stable/multimodal/clip_score.html

[4]https://github.com/openai/improved-gan

[5]https://github.com/w86763777/pytorch-image-generation-metrics/blob/master/pytorch_image_generation_metrics/inception.py

A. Qualitative Approach

Fig. 2a) exemplifies the representation of Capitu's eyes, often described as "*olhos de ressaca*" (sea surge eyes) in Machado de Assis's *Dom Casmurro*. These eyes are characterized as "gypsy eyes, oblique and dissimulated," as mentioned in chapters 13 and 32 of the book. The prompts used for this image were inspired by the physical and symbolic traits detailed in these chapters. Capitu is described as having brunette hair and light eyes, considering that the true color of her eyes is not explicitly revealed in the book. The prompt used was:

> "Painting of oblique and concealed eyes. Just eyes. Mysterious and energetic fluid, like the wave that retreats from the beach, on hangover days. Brunette, clear and large eyes, straight and long nose, had a thin mouth and wide chin."

Fig. 2b) clearly illustrates the importance of imagining a way to represent a scene. This image was created to depict Bentinho's suspicion that his son was fathered by Escobar, as described in chapters 131 and 132 of *Dom Casmurro*, portraying Bentinho's obsession with Escobar's photograph, which he kept in his office. Bentinho frequently noted the resemblance between his son and Escobar, fueling his suspicions. The image positions the child in front of Escobar's photograph, highlighting the similarities between them and symbolizing Bentinho's growing distrust. The prompt used was:

> "A painting of a young kid (4yo) in the center. In the background, there is a photograph of his father on the wall. The kid looks like the father, who was 40yo. 1800s."

Fig. 2c) portrays the opulence of the character *Ernestina*, based on *Júlia Lopes de Almeida's* novel *A Viúva Simões*. The image captures the essence of a rich widow, embodying her social status and physical attributes as described in chapter I of the book. The prompt used for this illustration was:

> "Concept art of a rich woman, widow, forty years old, bourgeoisie, a beautiful woman, tall, slender, beautiful black eyes, dark skin that was delicately feathery and soft, Rio de Janeiro. Digital artwork, illustrative, painterly, matte painting, highly detailed."

Fig. 2d) depicts *Aurélia* at the window with her suitors, as described in part 2, chapter 4 of *José de Alencar's* novel *Senhora*. Despite the prompt not specifying skin color, all 300 images generated for "a young woman, very beautiful" depicted white women, indicating a bias in the algorithm. The image captures the essence of *Aurélia's* character as she stands at the window, attracting numerous suitors. The prompt used for this illustration was:

> "Concept art of a young woman, very beautiful, stands in front of the window. She attracts a crowd of suitors who pass by in carriages and on foot. The eager looks and insinuating words of the suitors

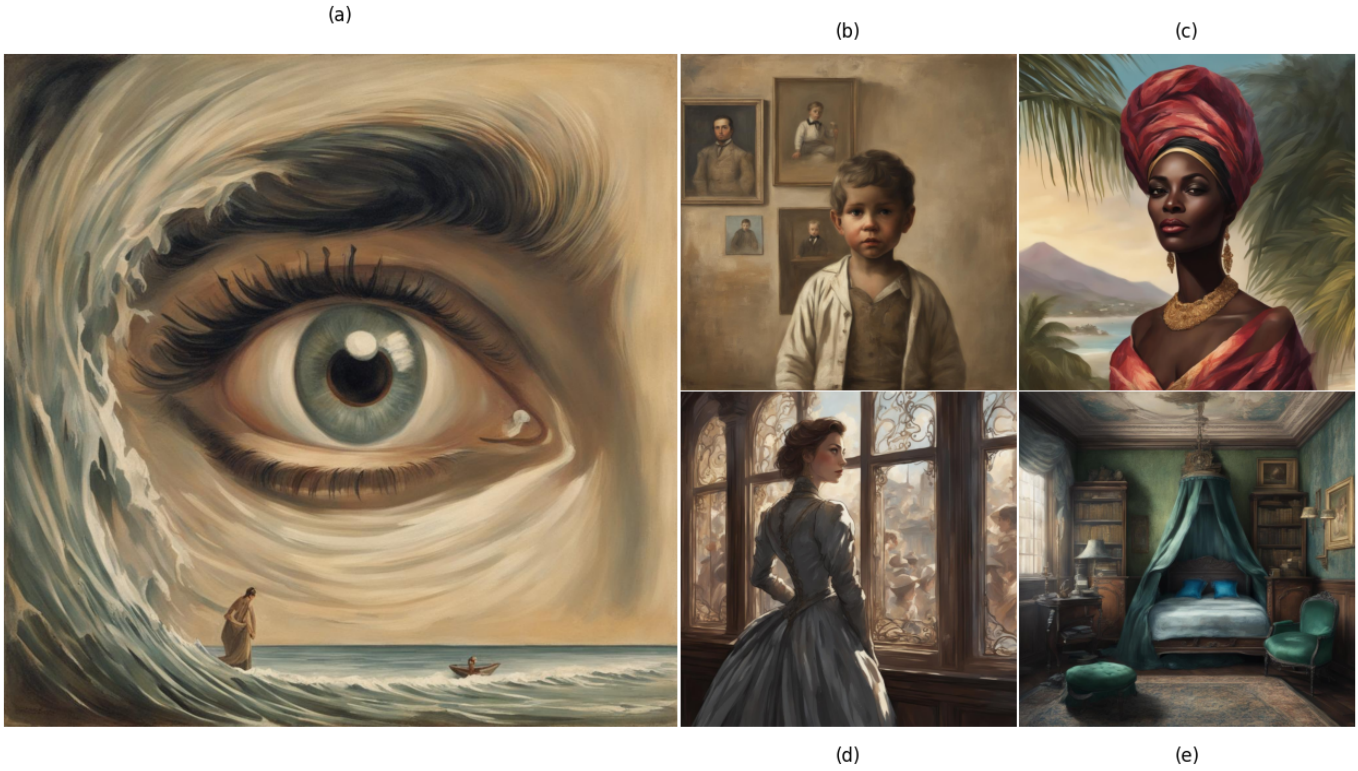


Fig. 2. Examples of generated images for (a, b) *Dom Casmurro*, (c) *A Viúva Simões*, and (d, e) *Senhora*.

> contrast with her cold impassivity; she remains at the window like a statue, fulfilling her duty, but without employing flirtatious tricks or seduction tactics."

Fig. 2e) illustrates *Seixas's* room, highlighting the stark contrast between his luxurious lifestyle and the poverty of his family, as described in part 1, chapter 5 of *José de Alencar's* novel *Senhora*. The scene and specifically the room was chosen for illustration due to the detailed descriptions in the book, which reveal aspects of *Seixas's* personality. This disparity is further emphasized as *Seixas* changes marriages twice, driven by his desire to escape his circumstances and secure the dowry offered to him. The generated images, however, did not fully capture all the objects mentioned in the prompt, seemingly focusing primarily on the books. The prompt used for this illustration was:

> "Concept art of modest, worn-out study with faded blue wallpaper, old furniture. Iron bed with green mosquito net contrasts with surroundings. Luxurious items like a tailored black coat, elegant evening wear, a Parisian hat, quality gloves, and fine boots seem out of place. The embroidered blue satin pillow stands out. The disorderly alcove with books, inkwells, ashtrays, and assorted trinkets contrasts with the well-appointed dresser counter. Corner with umbrellas, canes, some valuable, alongside artistic curiosities."

To address the issue of models "forgetting" details of characters or scenes in subsequent generations, it is recommended to use shorter and more concise prompts. Additionally, in our experiments, the model tends to prioritize sentences at the beginning of the prompt over those that follow, making it advisable to succinctly place the most important elements at the start. Reinforcing key elements by repeating the same token multiple times can also reduce the likelihood of certain details being "forgotten" in the prompt. As a final alternative, image-to-image techniques can be applied to the generated image to modify or add any missing traits, ensuring greater fidelity to the original context.

Fig. 3a) portrays a man from the Brazilian backlands, reflecting his determined nature and the harsh life he endures, as described in part 2, chapter III of *Euclides da Cunha's Os Sertões*. The prompt creation was based on: *"The sertanejo is, above all, a strong man. He does not have the exhaustive rickets of the neurasthenic mestizos of the coast. His appearance, however, at first glance, reveals the opposite. He lacks the impeccable physique, the poise, the correct structure of athletic organizations."*. Although the model faithfully represents the figure of the *sertanejo*, all 300 generated images show dense vegetation in the background, not reflecting the true scenery of the Brazilian backlands. The prompt used for this illustration was:

> "A painting of a man from the Brazilian backlands. He is a determined person, yet his countenance reflects the harsh life in the backlands, 19th century, realism."

Fig. 3b) illustrates a scene inspired by *Auta de Souza's* poem *"À minha avó"* from her book *Horto*. The generated image captures the warmth and comfort of the relationship between the grandmother and her grandson. The cozy environment, enhanced by elements such as a soft blanket, a cup of tea, and

Fig. 3. Examples of generated images for (a) *Os Sertões*, (b) *Horto*, (c) *A Viúva Simões*, (d) O cortiço, and (e) *O Triste Fim de Policarpo Quaresma*.

an old book, embodies the essence of comfort and security. However, it is important to note that the model also exhibited a bias towards generating white individuals, which is evident in the depiction of the characters. This particular book posed a challenge due to its abstract nature. The prompt used for the illustration was:

> "A painting of a grandmother and her grandson (9yr) sitting together in a cozy environment, surrounded by elements that symbolize comfort and security, a soft blanket, a cup of tea, and an old book, 19th century, realism."

Fig. 3c) illustrates the character *Luciano*, described in chapters I and II, from *Júlia Lopes de Almeida's A Viúva Simões*. The generated image portrays a man with gray hair, a thick build, consistent with the description provided. His expressive and friendly face, marked by dark circles under his eyes, captures the intended character traits. This image serves as a clear example of the model's tendency to create more stylized, cartoon-like illustrations when guided by prompts that include "concept art." The prompt used for this illustration was:

> "Concept art of a man, gray hair, thick, virile physiognomy, not slender, rounded belly, expressive and friendly face, dark circles under his eyes, Rio de Janeiro. Digital artwork, illustrative, painterly, matte painting, highly detailed."

Fig. 3d) depicts a scene from *Aluísio de Azevedo's* novel *O Cortiço* chapter XXI. This depiction highlights the model's ability to capture the specific details and atmosphere described in the prompt. The highly detailed, painterly style aligns with the intent to represent the character *João Romão*, a Portuguese owner of the collective housing, in a 19th-century Rio de Janeiro setting, showcasing the blend of period-appropriate elements with the everyday life of the character. The prompt used for this illustration was:

> "Concept art of a man walking back in flip-flops and a nightgown in the bedroom, wide room lined in blue and white with little yellow flowers pretending to be gold, a rug at the foot of the bed, and on the bench a nickel alarm clock, Rio de Janeiro, 19th century. Digital artwork, illustrative, painterly, matte painting, highly detailed."

Fig. 3e) presents the character *Policarpo Quaresma* from *Lima Barreto's O Triste Fim de Policarpo Quaresma*. The description comes from Chapter I and the image portrays a civil servant from the 19th century who is depicted as deeply valuing his country's culture. The illustration captures the essence of the character's dedication and connection to the cultural heritage of Rio de Janeiro during that period. The prompt used for this illustration was:

> "Concept art of a man, civil servant, values the country's culture, 19th century, Rio de Janeiro. Digital artwork, illustrative, painterly, matte painting, highly detailed."

Training generative TTI models relies on large datasets that can embed social and racial biases. Biased models in book illustrations can shape readers' perceptions, reflecting stereotypes instead of intended diversity, thus perpetuating harmful

narratives for marginalized communities. Broadly speaking, from an ethical standpoint, three primary issues highlighted in [19] are also immediately discernible in our work: misrepresentation (e.g., the harmful stereotyping of minority groups), underrepresentation (e.g., the eradication of the presence of a particular gender within specific professional roles), and overrepresentation (e.g., the predominance of Anglocentric viewpoints) [31].

B. Quantitative Approach

In this section, we present a quantitative analysis of the generated images using two evaluation metrics: CLIP and IS scores. The first metric evaluates the correlation between the input prompt and the generated image, with higher values indicating stronger semantic relevance between the text and image. The latter metric assesses both the quality and the diversity of the images produced by the generative model. Table I summarizes the evaluation results for each book concept.

TABLE I
EVALUATION RESULTS USING THE CLIP AND IS METRICS FOR EACH CONCEPT

Concept	CLIP (std)	IS (std)
A Viúva Simões - Júlia Lopes	20.44(1.63)	5.47(0.22)
Dom Casmurro - Machado de Assis	17.96(1.79)	5.91(0.32)
Horto - Auta de Souza	**21.05(2.30)**	6.35(0.34)
O Cortiço - Aluísio Azevedo	20.11(2.41)	3.76(0.23)
O Triste Fim de P. Quaresma - L. Barreto	18.94(2.23)	**7.87(0.36)**
Os Sertões - Euclides da Cunha	18.71(1.55)	6.20(0.45)
Senhora - José de Alencar	20.28(1.33)	3.86(0.23)

It should be observed that, despite the limited presentation of images — restricted to no more than two distinct prompts per book in Figs. 2 and 3 — each book is characterized by five distinct prompts with 300 images each. The aforementioned metrics are thus computed based on an aggregate of 1500 generated images. The results indicate varying levels of performance across different book concepts. For the CLIP Score, *Horto* by *Auta de Souza* achieved the highest score of 21.05 (standard deviation 2.30), reflecting strong semantic alignment between the generated images and the text prompts. This is followed by *A Viúva Simões* with a CLIP Score of 20.44 (standard deviation 1.63) and *Senhora* by *José de Alencar* with a score of 20.28 (standard deviation 1.33). On the lower end, *Dom Casmurro* scored 17.96 (standard deviation 1.79), indicating comparatively weaker semantic relevance.

In terms of the IS, *O Triste Fim de Policarpo Quaresma* achieved the highest score of 7.87 (standard deviation 0.36), suggesting superior image quality and variability. *Horto* follows with a score of 6.35 (standard deviation 0.34), indicating high-quality results but with less diversity compared to the previous. Conversely, *O Cortiço* and *Senhora* exhibited lower scores, 3.76 (standard deviation 0.23) and 3.86 (standard deviation 0.23) respectively, pointing to challenges in achieving high image quality and diversity.

Table I shows that books excelled in different metrics due to their unique textual traits. The CLIP Score likely favored Horto because its prompts were simpler and more descriptive, aligning closely with the text. O Triste Fim de Policarpo Quaresma had the highest IS, possibly due to its varied scene contexts, resulting in a diverse range of images. Books excel in different metrics because the metrics themselves evaluate distinct aspects of the generated images: CLIP focuses on semantic alignment between text and image, while IS measures the quality and diversity of the generated images.

IV. CONCLUSIONS

In this study, we explored the application of LDMs for the task of book illustration, utilizing the SD model to generate images based on prompts derived from seven classical Brazilian literary works. Our results indicate that the effectiveness of image generation is significantly influenced by the quality and specificity of the prompts provided. Prompts that were carefully crafted to encapsulate the essence of the scenes described in the books yielded more compelling and relevant visual results. Conversely, generic or poorly articulated prompts often led to suboptimal outcomes, highlighting the importance of prompt design in the generative process.

Through the illustrative examples provided, including figures representing characters and scenes from *Dom Casmurro*, *A Viúva Simões*, *Senhora*, *O Cortiço*, *O Triste Fim de Policarpo Quaresma*, and *Os Sertões*, we observed that the SD model could effectively capture and convey the thematic and visual elements of the source material. However, certain limitations were noted. For instance, the model's tendency to produce predominantly white individuals, despite the diversity described in the books, points to inherent biases in the training data that affect the generated results. This is exemplified by the generated images of characters from *Senhora* and *O Triste Fim de Policarpo Quaresma*.

Quantitative evaluations using CLIP and IS scores revealed variations in the quality of the generated images across different concepts. The analysis demonstrated that images aligned with literary descriptions from *Horto* and *O Triste Fim de Policarpo Quaresma* scored high in both metrics, indicating successful visualizations of these concepts.

Based on our findings, several key directions for future research emerge. First, refining generative models to reduce biases is essential. This can be achieved by retraining or fine-tuning models with more diverse datasets that better represent the cultural and historical contexts of the literary works, thereby enhancing the accuracy and inclusivity of the illustrations. The exploration of alternative diffusion models, as well as the integration of other methodologies such as GANs, could enhance output quality and mitigate inherent biases. Furthermore, conducting an ablation study, such as removing components like the VAE and evaluating the impact on the illustrations using quantitative metrics, would provide valuable insights into the contributions of different model components. This approach would help optimize the generative process by identifying which elements are most critical for improving accuracy and aesthetics. The development of new evaluation metrics that consider cultural relevance, emotional tone, and symbolic representation—supplemented by human evaluators

such as literary scholars and art experts—is also important to produce nuanced assessments and achieve better alignment between text and image. Finally, collaborative approaches with literary scholars in the illustration process can also ensure that the generated illustrations faithfully reflect the source material, thereby resonating more deeply with the intended audience.

ACKNOWLEDGMENTS

F. M. acknowledges support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), project number 88887.607339/2021-00. A. F. Z. acknowledges support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), project number 140935/2024-0.

REFERENCES

[1] OpenAI, "Gpt-4 technical report," 2023, doi: https://doi.org/10.48550/arXiv.2303.08774.

[2] K. I. Roumeliotis and N. D. Tselikas, "Chatgpt and open-ai models: A preliminary review," *Future Internet*, vol. 15, no. 6, p. 192, 2023, doi: https://doi.org/10.3390/fi15060192.

[3] J. Huang and M. Tan, "The role of chatgpt in scientific communication: writing better scientific review articles," *American Journal of Cancer Research*, vol. 13, no. 4, p. 1148, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37168339/

[4] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017, doi: https://doi.org/10.48550/arXiv.1703.10135.

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016, doi: https://doi.org/10.48550/arXiv.1609.03499.

[6] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019, doi: https://doi.org/10.3390/app9194050.

[7] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264403242

[8] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021, doi: https://doi.org/10.48550/arXiv.2102.12092.

[9] (2022) Midjourney. [Online]. Available: https://www.midjourney.com

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022, doi:https://doi.org/10.48550/arXiv.2112.10752.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[12] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4066–4079, 2018, doi:https://doi.org/10.1109/TIP.2018.2836316.

[13] J. Liu, C. Wang, H. Su, B. Du, and D. Tao, "Multistage gan for fabric defect detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3388–3400, 2019, doi:https://doi.org/10.1109/TIP.2019.2959741.

[14] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505–1514, doi:https://doi.org/10.1109/CVPR.2019.00160.

[15] R. S. Casado and E. C. Pedrino, "A comparison study of depth map estimation in indoor environments using pix2pix and cyclegan," *IEEE Latin America Transactions*, vol. 22, no. 3, pp. 213–221, 2024, doi: https://doi.org/10.1109/TLA.2024.10431422.

[16] K. Walczak and W. Cellary, "Challenges for higher education in the era of widespread access to generative ai," *Economics and Business Review*, vol. 9, no. 2, pp. 71–100, 2023, doi:https://doi.org/10.18559/ebr.2023.2.743.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020, doi:https://doi.org/10.48550/arXiv.2006.11239.

[18] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, doi:https://doi.org/10.1109/TPAMI.2023.3261988.

[19] H. Vartiainen and M. Tedre, "Using artificial intelligence in craft education: crafting with text-to-image generative models," *Digital Creativity*, vol. 34, no. 1, pp. 1–21, 2023, doi: https://doi.org/10.1080/14626268.2023.2174557.

[20] A. R. Doshi and O. Hauser, "Generative artificial intelligence enhances creativity," *Available at SSRN*, 2023, doi: http://dx.doi.org/10.2139/ssrn.4535536.

[21] H. Smart, "Making books with generative ai," Master's thesis, Concordia University, December 2023, unpublished. [Online]. Available: https://spectrum.library.concordia.ca/id/eprint/993284/

[22] B. Tomlinson, R. W. Black, D. J. Patterson, and A. W. Torrance, "The carbon emissions of writing and illustrating are lower for ai than for humans," *Scientific Reports*, vol. 14, no. 1, p. 3732, 2024, doi: https://doi.org/10.1038/s41598-024-54271-x.

[23] W. A. C. Castañeda, M. Amadeus, A. F. Zanella, and F. R. P. Mahlow, "Generative ai methodology for producing assisted art: Representation of the historical-cultural identity of southern brazil," in *Making Art With Generative AI Tools*. IGI Global, 2024, pp. 179–196, doi: https://doi.org/10.4018/979-8-3693-1950-5.ch010.

[24] A. M. Piskopani, A. Chamberlain, and C. Ten Holter, "Responsible ai and the arts: The ethical and legal implications of ai in the arts and creative industries," in *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, ser. TAS '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3597512.3597528

[25] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," 2023. [Online]. Available: https://arxiv.org/abs/2307.01952

[26] G. Ilharco, M. Wortsman, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5143773

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

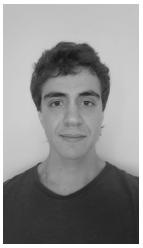
[28] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023, doi:https://doi.org/10.48550/arXiv.2307.01952.

[29] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," 2022, doi: https://doi.org/10.48550/arXiv.2104.08718.

[30] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," 2016, doi: https://doi.org/10.48550/arXiv.1606.03498.

[31] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021, doi: https://doi.org/10.48550/arXiv.2108.07258.

Felipe Rodrigues Perche Mahlow received a teaching degree in Physics and a bachelor's degree in Materials Physics from São Paulo State University (Unesp), in 2020 and 2023, respectively. He is currently pursuing a Ph.D. degree in Computer Science at São Paulo State University (Unesp), with a focus on classical and Quantum Machine Learning and their applications to Quantum Information Science. His research interests include Artificial Intelligence and Quantum Computing.



André Felipe Zanella received a Bachelor's degree in Mathematics and a Master's degree in Computer Science from the State University of Maringá, in 2022. He is currently pursuing a Ph.D. in Computer Science at the same institution. His research focuses on optimization, Machine Learning, and TTI Diffusion Models.

William Alberto Cruz Castañeda received a Bachelor's degree in Computer Science from Benemérita Universidad Autónoma de Puebla and a Bachelor's degree in Computer Engineering from the Federal University of Rio Grande do Sul. He holds Master's and Ph.D. degrees in Electrical Engineering, with a focus on Biomedical Engineering, from the Federal University of Santa Catarina, and a Postdoctoral degree in AI and Biomedical Engineering from the State University of Santa Catarina. He is currently a professor at the Federal Technological University of Paraná. His research interests include Ubiquitous Computing, Machine Learning, and GenAI.

Regilene Aparecida Sarzi-Ribeiro received Postdoctoral degrees in Poetics and Cultures in Digital Humanities from UFG (2022) and Arts from UNESP/SP (2013). She holds a Ph.D. in Communication and Semiotics from PUC/SP (2012), a Master's in Arts from UNESP/SP (2007), and a degree in Art Education with a specialization in Visual Arts from FAAC - UNESP, Bauru/SP (1994). She is a faculty member at UNESP/Bauru, where she coordinates the Graduate Program in Media and Technology. She is also part of the Chair of Design, Art, and Science at Media Lab BR and a research member of Red de Investigación de la Imagen at the University of Málaga, Spain.