CINTED-UFRGS Revista Novas Tecnologias na Educação

Sprites for educational games: an empirical study with generative artificial models

Fernanda Pires, ThinkTEd – UEA, fpires@uea.edu.br, 0000-0002-8180-9850
Fabrizio Honda, ThinkTEd – UEA, fabrizio.honda@icomp.ufam.edu.br, 0009-0003-8174-1002
Jeniffer Macena, ThinkTEd – UEA, jeniffer.souza@icomp.ufam.edu.br, 0009-0004-7163-629X
Marcela Pessoa, ThinkTEd – UEA, mpessoa@uea.edu.br, 0000-0004-3654-5585

Abstract: Creating sprites for educational games presents challenges for computing teams related to artistic competence and the scarcity of suitable free resources in the educational field. This study examined the application of Generative Artificial Intelligence (GAI) models in generating sprites for educational games. We presented the research as an empirical, qualitative study with an exploratory design. The models generated characters (sprites) based on prompts we developed, and educational game experts evaluated them. The analyses include descriptive statistics and thematic analysis. The results reveal that all models generated adequate sprites, with DALL-E standing out in terms of average rating. The qualitative analysis highlighted the models' potential for prototyping, despite ethical concerns and technical limitations.

Palavras-chave: Generative Artificial Intelligence, Sprites, Educational Games, Prototyping, Qualitative Evaluation

Sprites para jogos educacionais: um estudo empírico com modelos artificiais generativos

Resumo: A criação de sprites para jogos educacionais apresenta desafios para equipes de computação, relacionados à competência artística e à escassez de recursos gratuitos adequados na área educacional. Este estudo examinou a aplicação de modelos de Inteligência Artificial Generativa (IAG) na geração de sprites para jogos educativos. Apresentamos a pesquisa como um estudo empírico e qualitativo com delineamento exploratório. Os modelos geraram personagens (sprites) com base em prompts que desenvolvemos, e especialistas em jogos educativos os avaliaram. As análises incluem estatística descritiva e análise temática. Os resultados revelam que todos os modelos geraram sprites adequados, com o modelo DALL-E destacando-se em termos de avaliação média. A análise qualitativa evidenciou o potencial dos modelos para prototipação, apesar das preocupações éticas e das limitações técnicas.

Keywords: Inteligência Artificial Generativa, Sprites, Jogos Educacionais, Prototipação, Avaliação Qualitativa

1. Introduction

Educational games have gained prominence as a tool to support learning, allowing the incorporation of curricular content in a fun and playful way, creating rich learning environments capable of engaging students on cognitive, affective, and behavioral levels (Plass; Homer e Kinzer, 2015). In these games, character design plays a fundamental role that is not merely aesthetic: it contributes to players' motivation, engagement, and empathy (Klopfer *et al.*, 2025), as observed in some studies (Rativa; Postma e Zaanen, 2020; Carvalho, 2024; Ho e Ng, 2022). Thus, the appearance of characters, including expressiveness, personality, and visual coherence, significantly influences the emotional attachment that users can develop towards the game (Bopp *et al.*, 2019).

However, creating sprites for games presents several obstacles, including the time required to develop the art, expertise in character illustration, the lack of free materials consistent with the game's theme, and a shortage of sufficient staff, among others (Grassi, 2021; Battistella; Wangenheim e Fernandes, 2014). Furthermore, in the case of educational games, game creators need to consider other factors when designing sprites, such as: cognitive load, to avoid overloading the player with unnecessary elements (Das *et al.*, 2024); the relationship with the curriculum content, to prevent disconnecting it from the learning mechanics (Boller e Kapp, 2018); and avoiding elements that incite violence, such as weapons and blood (Asadzadeh *et al.*, 2024); among others.

One alternative that can mitigate these challenges is the use of Generative Artificial Intelligence (GAI), particularly when combined with diffusion models and Large Language Models (LLMs). In this process, a user provides input (prompt), whose LLM interprets the textual information and generates semantic representations, guiding the diffusion model to create multimodal content (such as images). The diffusion model then performs visual synthesis and returns a coherent and contextually relevant result (Benjdira *et al.*, 2025). Because of their capabilities in generating high-quality illustrations, the integration of diffusion models with LLMs has fostered research in the generation of sprites, scenes, and other visual elements for games (Wu *et al.*, 2025).

This study was motivated by a challenge identified in an educational game improvement project at the laboratory ThinkTEd[*]. The project involved refactoring games previously evaluated in terms of game design, incorporating suggested improvements, such as redesigning levels and creating new visual art to enhance the games' appeal to the public. The team, composed mostly of computer science students, lacked designers and illustrators. Faced with these limited specialized resources, the students turned to free sprite platforms, such as *CraftPix*[†].

Thus, considering the challenge in the refactoring project at ThinkTEd, the importance of characters for player motivation and engagement, the difficulty of game designers in creating or localizing sprites, and the potential of diffusion models integrated with LLMs, this work presents the following research question (RQ): "How do diffusion models integrated with LLMs perform in the creation of sprites for educational games?". We conducted an empirical study to answer the RQ.

Although the RQ discusses how diffusion models behave with LLMs in sprite creation, this study focuses only on how educational game developers view the visual appearance of the generated sprites. Teachers and university students with experience in educational games evaluated the sprites; they did not use the models directly but analyzed previously generated sprites based on visual criteria and how well they fit the academic context. Therefore, the conclusions pertain only to their aesthetic perception and the visual applicability of the sprites in educational games.

2. Foundations and related work

Generative Artificial Intelligence (GAI) consists of the application of generative modeling and deep learning to understand and/or generate text, graphics, audio, and video (Jovanovic e Campbell, 2022; Sengar *et al.*, 2024). The emergence of Large Language Models (LLMs) recently boosted this field: models that generate text from a given input (prompt), based on a network architecture with attention mechanisms. Given their capabilities, researchers have applied LLMs in conjunction with other models, enabling advances in Vision-Language Models (VLMs): multimodal models that combine visual and textual information, focused on activities such as captioning images, generating images from textual descriptions, and responding to visual queries (Ghosh *et al.*, 2024).

Parallel to the emergence of LLMs, the use of diffusion models has also grown

[*] Available at: https://thinktedlab.org/

[†] CRAFTPIX. Home page. 2025. Available at: ⟨https://craftpix.net/⟩

(Chen *et al.*, 2025). During training, these models learn to degrade authentic images by adding progressive noise, gradually corrupting them until they become unrecognizable. Then, in the generation process, they apply the reverse process: starting with a noisy image and progressively removing the noise to restore it. In this way, they enable the generation of realistic images, especially text-driven ones. Studies investigate their use for various applications, such as sprite (character) generation for games (Hsieh; Zhang e Yan, 2024; Merino *et al.*, 2023). Furthermore, with the advancement of the AI field, LLMs have been utilized in conjunction with diffusion models, facilitating the interpretation of semantic data that guides image generation. In this regard, we describe works found in the literature that utilize diffusion models (integrated or not with LLMs) to generate sprites for games, as outlined below.

The work by Hsieh et al. (2025) proposes “Sprite Sheet Diffusion”, a method based on diffusion models for generating animation sequences (sprites) from a reference image and target poses. The authors constructed a dataset with 152 action sequences and adapted the Animate Anyone model for this task. They then compared the method with adjusted and unadjusted versions of SD-IPCN and Animate Anyone, which achieved better quality and consistency. Thus, one of the main contributions of the work is the proposal of an approach that reduces the manual effort in creating animations.

In Gallotta et al. (2025), the authors investigate the use of diffusion models to generate context-adapted game sprites using LLMaker, which combines Stable Diffusion with LLM GPT-3.5-Turbo. They tested the tool on 50,000 sprites with varying levels of context, including colors, descriptions, and backgrounds, and evaluated it using both automatic image metrics and a study involving 16 participants. The results demonstrated that contextual information improves the visual coherence of sprites, facilitating the customization and automation of these elements.

The research by Dai et al. (2024) proposes a method based on diffusion models capable of generating game levels (procedurally) from a single example. The authors tested the model on games such as Minecraft and Super Mario Bros., using dense token representations and a latent diffusion network with a restricted receptive field. The results demonstrate that the model outperforms GAN-based approaches in terms of level coherence and diversity. Although it does not involve sprites, the study contributes to the automatic generation of visual elements in games, such as levels and scenarios.

Table 1 presents a comparison of related works with this article, showing that only Gallotta et al. (2025) use diffusion models integrated with LLMs, while the others focus on other approaches. All works focus on generating visual elements for games, especially sprites. The innovative aspects of this work include: (i) the comparison of four diffusion models, which two has integration with LLMs; (ii) the focus on generating sprites for educational games, considering learning aspects; (iii) the implementation of prompt tests, which can assist researchers in the reproducibility of the study; and (iv) the evaluation of the sprites by educational game developers, who already have experience producing/localizing sprites appropriate to the theme of their games.

Table 1. Comparison of work characteristics.

Work	**LLMs**	**Diffusion Models**	**Game Visual Elements**	**Educational Game**
Hsieh et al. 2024	-	x	x	-
Gallotta et al. 2025	x	x	x	-
Dai et al. 2024	-	x	x	-
This Work	x	x	x	x

3. Materials and Methods

This section presents the scientific and methodological procedures we adopted in this research.

3.1. Study design

This work describes a qualitative, empirical study with an exploratory design (Creswell e Poth, 2016). Its objective is to analyze the capacity of Generative Artificial Intelligence (GAI) models, namely diffusion models and LLMs, in the creation of sprites for educational games, considering the practical applicability and the perceived quality of the artifacts generated by creators and digital educational games. Computer science students from the Computing Education program at Amazon State University (UEA) executed the research, investigating the production and use of digital educational games as learning tools. However, the team does not include a design or art professional.

3.2. Procedures

The procedures for conducting this study are outlined in this subsection, comprising four main steps, as described below.

Model selection: For this investigation, we used four prompt-based image generation models: DALL-E 3[‡], Krea.AI 1[§], Imagine.Art 1.0 Lite[¶], and Leonardo AI Flux Dev[‖]. We accessed each tool in its free version, considering its default settings at the time of use.

Selection of educational game for sprite testing: We selected the game"Gramágica" (Macena *et al.*, 2019), which focuses on practice syllabic classification. The game designers developed the game mechanics and gameplay based on the pillars of Computational Thinking. The motivation for choosing it was: (i) the game had related scientific publications in the literature; (ii) it contained a demonstration video on YouTube; and (iii) it provided access to the game. In the game's story, a young sorceress resides in a kingdom where a magical jewel has the power to organize words correctly. A powerful sorceress guards this jewel with a wheel; however, the jealous wizard Rusbel decides to take the jewel by force, imprisoning the guardian in her tower. Faced with this threat, Granger embarks on a challenging journey to rescue the sorceress and restore linguistic balance to the kingdom. Each level of the game corresponds to a syllabic category. For example, in Level 1 (Figure 1(a)), the character begins her journey in search of monosyllabic words (Figure 1(b)). By correctly collecting the proposed words, the player unlocks the next level. The documentation is available in supplementary material.[**]

Prompt design: After selecting the educational game, we chose two characters for image generation studies: Granger, the heroine, and Rusbel, the villain. We carried out the

[‡]https://openai.com/dall-e
[§]https://www.krea.ai
[¶]https://www.imagine.art
[‖]https://leonardo.ai
[**]Game Design Document (GDD) of "Gramágica" ⟨https://drive.google.com/drive/folders/1RMxZ2KXkFRftbrkV_9ZtfkANMRzYOG58?usp=sharing⟩

(a) Monosyllabic words level screen. (b) Disyllabic words level screen.

Figure 1. Game screens from levels 1 and 2, respectively.

prompt design as follows: (i) we sent the original sprites of these characters to ChatGPT, along with the generation criteria of a consistent 2D, canvas, and chibi style, to present a user-friendly look appropriate for the target audience. Specifically, for the Leonardo AI and Imagine.art models, we have two interaction fields: one with positive prompts (what to look for in the generated output) and another with negative prompts (what to avoid in the generated output). Therefore, we generated additional prompts for these models to address these requests; (ii) ChatGPT created a prompt describing these sprites, considering the informed criteria; (iii) we tested the prompts on the diffusion models selected in the "*Model Selection*" phase; (iv) before generating the images, we refined the prompts in several interactions with the models so that they met aesthetic coherence with the "Gramágica" look, resulting in the prompts available at the following link.[††]

Prompt application: We sent the prompts to the previously selected models, resulting in the sprites in Figure 2. Each model presented a particularity in the generation process: (i) in DALL-E, it was necessary to insert and send the previously created prompt. After that, the model generated the sprite according to the description provided; (ii) in Leonardo AI, we chose the base model "Concept Art" , we inserted the negative and positive prompts constructed in the previous step, we selected the visual style "Game Concept Cartoon", and we set the contrast to "Medium". Then, the model generated four sprite options, from which we selected the one that best corresponded to the character; (iii) in Imagine.art, we chose the base model "Anime 2.0" (the one that best suited the sprite context), we inserted the previously developed positive and negative prompts, we set the style to "Cartoon", and we selected the camera perspective to "Wide Shot". The model then generated the sprite ;[‡‡]and (iv) in Krea.ai, we inserted the previously constructed single prompt, and the model generated the sprite. The free version of the model does not allow for advanced style/template selection, limiting customization.

3.3. Evaluation

To conduct the evaluation stage of the research, we developed an evaluation questionnaire to assess the suitability of the sprites. The questionnaire begins with a section detailing the research objective, the types of questions, and information about the educational game (Gramágica). It also includes a consent form, allowing evaluators to determine whether or not they agree to the anonymous disclosure of data for research purposes only. If they agree, they proceed to the following four sections of the questionnaire: (i) general information – information about the evaluators' profiles; (ii) educational games – years of experience with educational games, number of games created, method they used to obtain sprites, and difficulty of this process; (iii) and

[††]The prompts are avaliable at:⟨https://drive.google.com/drive/folders/1gaoQo3dNesrlIt0itYyDTfnS51mH6l6q⟩.
[‡‡]With the Premium version, it is possible to generate up to four images at a time, but the free version is limited to only one.

Figure 2. Sprites of the characters Granger and Rusbel generated by the models.

(iv) address, respectively, the evaluation of the sprites of the characters "Granger" and "Rusbel", containing the figures of the sprites which each of the four models generated, with five associated quantitative questions on a Likert-5 scale, regarding alignment with the game's narrative, pleasant visuals, adequate representation, appropriate colors, and coherence with the game's visual style. Additionally, there is a qualitative (open-ended) question about what the evaluators thought and what the positive and negative points were. At the end of both sections, there are questions to select the model that generated the best sprite and another to justify the selection. At the end of the form, there is a question to assess whether participants agree that the diffusion models and LLMs are capable of generating sprites for educational games, as well as a qualitative question about their thoughts on this process. We emphasize that we anonymize the names of the LLMs (referred to as Model 1, Model 2, among others) to avoid bias in the evaluation.

Since the goal of the study was to evaluate whether generative AI models created suitable sprites for a specific educational game, we looked for people with experience in developing educational games (developers and/or artists). To ensure greater methodological rigor and include a wide range of perspectives in the study, we chose to conduct a broad survey by sharing the questionnaire with university teachers and students nationwide. The only requirement we set was previous direct or indirect involvement in the development of educational games. Based on the collected data (evaluations in the questionnaire), we perfomed the following analyses: (i) visual, using a boxplot to illustrate the evaluations of each AI model and a heatmap matrix with the sum of the models' scores for each question; (ii) statistical, using descriptive statistics with mode, median, minimum and maximum values, standard deviation, range, and coefficient of variation; and (iii) thematic analysis for the qualitative questions, which involved an initial reading of the responses to define thematic categories. Next, we grouped the responses according to their respective categories to observe patterns and extract insights.

4. Results and discussion

Regarding the evaluation questionnaires, a total of 30 people agreed to participate in the survey. All participants provided informed consent within the questionnaire, being aware that their participation was voluntary and that responses would be recorded anonymously for academic purposes. The profiles were as follows: (i) university – 57% belonged to Amazon State University (UEA), 27% to Federal Amazon University (UFAM), and 16% to none or other universities; (ii) course – 90% were linked to

The participants had access to the game, both the application and its documentation undergraduate or postgraduate courses in computing; (iii) experience with educational games – 40% had 1 to 3 years of experience, 27% had less than one year, 20% had 3 years or more, and 13% had none; (iv) creation of educational games – 50% participated in one or two, 27% in none, and 23% in three or more; (v) method of obtaining sprites – the majority (87%) use ready-made sprites from a website (CraftPix, GameArt 2D, etc.), 23% ask a friend/colleague to draw them, 20% draw them themselves, and 27% use compositing (cutting assets from various websites); and (vi) difficulties in this process – access to free sprites, difficulty in finding sprites suitable for the educational content and the game, limitations in design or creative expertise, limited resources (such as full animations), among others. After completing the form, we stored the data in an online spreadsheet.

Regarding the participants' evaluations of the sprites generated by the diffusion models with LLMs, the boxplot in Figure 3 shows the distribution of the sum of scores assigned by each participant, organized by character and model. The X-axis corresponds to the character evaluated (Granger or Rusbel), and the Y-axis indicates the range of the sum of scores, ranging from 5 (the evaluator assigned a score of 1 to all questions) to 25 (the maximum score assigned by the evaluator in all questions).

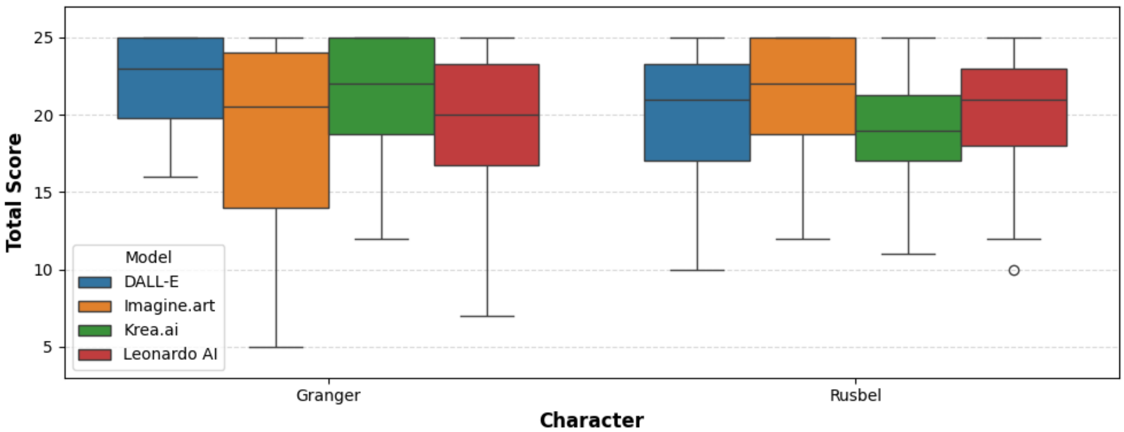

Figure 3. Boxplot graph showing the distribution of model evaluations by character.

Overall, the evaluations for each model were similar. For the character "Granger", the DALL-E and Krea.ai models received the best evaluations, with the highest sums and the lowest data dispersion. The Imagine.art model, although receiving some high scores, presented the most significant variability in responses, reflecting a divergence among participants regarding its quality. Leonardo AI, on the other hand, received a middling evaluation, with both high and low scores. For the character "Rusbel", the evaluations were more varied: Imagine.art, despite not having the lowest dispersion, received the most positive evaluations. Krea.ai, which received a good review for the previous character, received the lowest result compared to the other models, despite having the weakest data dispersion. Leonardo AI presented a more concise distribution, with mostly positive evaluations. DALL-E remained consistent with positive evaluations but presented a greater number of low assessments.

To investigate the performance of the models on each question, we created a heatmap graph (Figure 4). The X-axis represents the questions in the questionnaire, and the Y-axis represents the names of the models evaluated (M1 – DALL-E; M2 – Imagine.art; M3 – Krea.ai; M4 – Leonardo AI). The cells contain the sum of the scores assigned by the evaluators to the questions related to each character. Thus, the scores range from 60 (if all 30 evaluators assigned a score of 1 on the question for both sprites in the model) to 300 (the maximum score assigned by all evaluators).

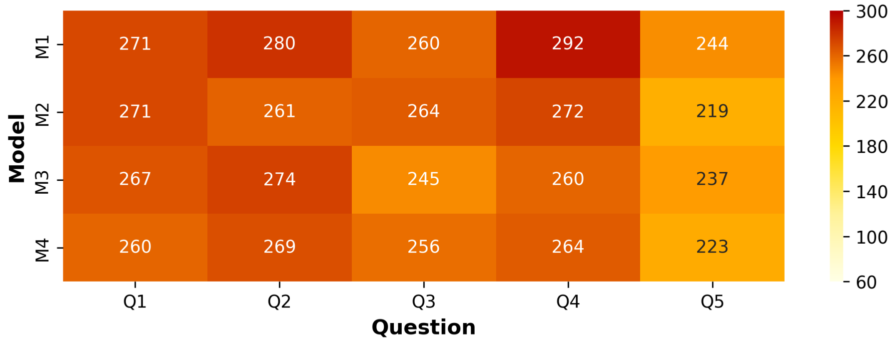

Figure 4. Heatmap graph with the sum of question scores for each model.

Regarding Q1 (alignment with the game's narrative), all models presented satisfactory and similar results, indicating that the generation was faithful to the provided prompt. For Q2 (pleasing visuals), the responses were also positive, highlighting the models' ability to generate aesthetically appropriate sprites, especially the DALL-E (M1) and Krea.ai (M3) models. Regarding Q3 (adequate character representation), the scores showed a slight decrease but remained high, indicating that the models effectively conveyed the essence of the characters. Q4 refers to appropriate colors, whose scores were positive, especially for DALL-E. This point reinforces that the models accurately interpreted the prompt. Finally, Q5 (consistency with the game's visual style) received the lowest scores, although they are above the average for the range (180). This point corroborates the evaluators' divergence regarding the sprites' adherence to the Gramágica style.

For a more detailed analysis, we created Table 2 with the descriptive statistics for each model, unifying the evaluations of both sprites, including their medians, standard deviations (SD), minimum and maximum values, amplitude, and coefficient of variation (CV). Analysis of the table shows that evaluators considered DALL-E (M1) the most satisfactory model, with the best median value (271), minimum value (244), and maximum value (292). This point may be related to the model's architecture, which integrates a diffusion model with an LLM. In contrast, Imagine.art (M2) obtained the lowest scores, presenting the highest standard deviation (21.96), the lowest minimum value (219), a high amplitude (53), and the highest coefficient of variation (8.53). One of the styles defined for this model was "Anime 2.0", whose evaluations suggest that, although it is a suitable sprite, it is inconsistent with the art style for educational games. As for Krea.ai (M3), the results were positive: a lower standard deviation (15.32), satisfactory minimum and maximum scores, and a lower range (37) and coefficient of variation (5.98). Although its median was one of the lowest, it was similar to the other models. Its overall performance was as good as DALL-E, despite the limitation of its free version, which does not allow the selection of advanced styles or models. On the other hand, Leonardo AI (M4) presented the lowest median (260) and maximum value (269). Despite this, its values were not low, presenting evaluations similar to the other models. We assumed that one of the criteria that lowered the model's score was the style of the character "Rusbel", which had a more three-dimensional appearance (not suitable for "Gramágica").

Regarding which model the evaluators considered to have generated the best sprite, 14 (47%) chose DALL-E for the character "Granger". Most of the justifications

Table 2. Descriptive Statistics of Model Evaluations

Model	Median	SD	Minimum	Maximum	Range	CV (%)
M1	271.0	18.43	244.0	292.0	48.0	6.84
M2	264.0	21.96	219.0	272.0	53.0	8.53
M3	260.0	15.34	237.0	274.0	37.0	5.98
M4	260.0	18.20	223.0	269.0	46.0	7.15

referred to coherence with the game's visual style, design aspects (lines, colors, and minimal shadows), simplicity, suitability for the audience, ease of use (according to P17, developers can already import the sprite into the Unity game engine), ease of customization due to the type of art, conformity with the game's demo video, among others. Ten people (33%) opted for the sprite generated by the Leonardo AI model, highlighting aspects such as: fidelity and conformity with the character's description and requirements, aesthetic quality, fidelity to the game's narrative and purpose, suitability for the audience, among others.

Regarding the character "Rusbel", the DALL-E and Leonardo AI models were also the most predominant from the evaluators' perspective, with 43% and 33% preference, respectively. Regarding DALL-E, the evaluators also highlighted its consistency with the game's visual style and the character's description. P19 points out that the other sprites "disagree in many ways as the detail of the generated sprites increases", highlighting the sprite's simplicity. Moreover, despite its simplicity, P25 comments that "it had better contrast and use of visual elements to match the character's defined personality". Regarding Leonardo AI, the positive points include its suitability for the narrative, consistency with the character's description, and visual style. Furthermore, as P22 and P24 indicated, the sprite more closely resembles a villain.

A curious factor when analyzing the evaluators' preferences concerns the Leonardo AI model. Although it received the lowest quantitative ratings, evaluators considered it the second-best sprite-generating model for both "Granger" and "Rusbel", with 33% preference for each (10 out of 30). This contrast indicates that, despite the lower objective ratings, the reviewers reflected on it and valued it more highly when comparing the sprites with those of the other models. In contrast, Krea.ai, which received satisfactory ratings and the lowest standard deviation, was the evaluators' least preferred model. This point indicates that, despite its consistency and suitability for the game's style, the sprites from other models were considered more suitable.

Regarding the question "What do you think about using diffusion models and LLMs to generate sprites for educational games?", we perfored a thematic analysis. We classified the evaluators' responses to the question into five analysis categories: (i) **full approval** without reservations (60%, n=18), highlighting the usefulness of using models in small teams or without illustrators, the ease in creative processes and prototyping, an alternative where there are few free assets; (ii) **partial approval with technical limitations** (20%, n=6), in this group they report the difficulty in maintaining standards, pointing out as needs: having a well-described prompt, the challenges in creating animations, among others; (iii) **approval with ethical reservations** (10%, n=3) in this group, people indicate the use of models as a valid alternative, but warn of ethical issues, mainly related to copyright (appropriation of styles); (iv) **total rejection** (6,7%, n=2) – they disagree with the use of models, with arguments such as the loss of the game's "magic" (as pointed out by P14); and (v) not applicable (3,3%, n=1) – answers that disagree with the question. Thus, it is possible to observe that the use of GAI to generate images for digital games is a viable option. We expected reservations, considering the novelty factor and the numerous discussions in the ethical and educational fields. However, for producers of educational games who do not have large teams and in production scenarios, especially within universities, it can be a valid alternative.

Thus, answering the research question, the results suggest that no model is predominantly superior to the others, but rather that each has distinct characteristics, and its applicability depends on the context and the subjective perspective of the evaluators. All models demonstrated the ability to generate appropriate sprites according to the prompt and presented similar scores for both characters in the game “Gramágica”. To achieve this goal, it was necessary to follow a detailed step-by-step process, including the following steps: defining the criteria for model selection, selecting the educational game and target characters, developing the prompt, generating the images, developing the evaluation instrument, and collecting/analyzing the data.

Limitations of this study include: (i) most participants are experts in educational games and lack illustration expertise. In future research, we plan to involve illustrators in the evaluation process; (ii) the models only generated one frame for each sprite. The ideal is to present a set of sprites to simulate an animation and investigate the models' ability to maintain visual consistency, which is also the focus of the following study; (iii) evaluation biases, since the sprites refer to an existing game. This point may have biased participants' evaluations toward models more similar to the original sprites. We will address this point in subsequent studies, in which a character will be created from a game still in development; and (iv) the choice of model and prompt styles may not have been the most suitable. We intend to address this issue in the future by incorporating an analytical process, examining the specific characteristics of the models, and refining the prompts.

5. Conclusions

Given the importance of developing educational games with elements that make their content more playful and engaging for the target audience, it is essential to ensure the quality of the materials used in their interfaces. This quality must consider mechanisms that explore information processing channels, as proposed by Mayer (2009). Creating games is a complex process, as it involves multiple areas, and this challenge intensifies when it demands the expertise of the teams involved, especially in terms of illustration/design skills. Furthermore, locating free assets that are suitable for game design is challenging. In this regard, the use of diffusion models and LLMs has intensified, focusing on the artificial generation of these sprites. Considering this context, this work conducted an empirical study to answer the following RQ: “How do LLMs and diffusion models perform in the creation of sprites for educational games?”.

To answer this question, we conducted the following steps: (i) Model selection – we chose four models through benchmarking (DALL-E, Leonardo AI, Imagine.art, and Krea.ai); (ii) Game selection – we selected the educational game “Gramágica”; (iii) Prompt design – we created and refined textual prompts based on the original sprites of the characters “Granger” and “Rusbel”, considering the 2D, canvas, and chibi styles; (iv) Prompt application – we applied the prompts to the models, generating the sprites according to the specific configurations of each platform; (v) Evaluation – we developed a form with qualitative and quantitative approaches, including the game context and research objectives; (vi) Data collection – 30 participants, with experience in educational games and computer science degrees (undergraduate and graduate), responded to the questionnaire; and (vii) Data analysis – we performed visual analyses, descriptive statistics, and thematic analysis to interpret the collected data.

The results indicate that all models were able to generate sprites appropriate to the prompts, although with varying performance for each character and the evaluated criteria. The DALL-E stood out with the highest averages and was the evaluator's most preferred model, mainly due to its fidelity to the game's style. Despite receiving the lowest averages in the quantitative evaluations, the Leonardo AI model was the second most recommended, suggesting that factors such as fidelity to the narrative and style positively influenced its perception. Finally, the Krea.ai model received consistent evaluations but was the least cited as a favorite, and Imagine.art showed the most significant variation in scores, with poorer performance in adherence to the educational style. Thus, it is clear that the model depends on the context, the technical criteria, and the level of experience and familiarity of the evaluators.

In the field of study involving LLMs and diffusion models, there is a lack of transparency regarding the training data used for these models. In other words, the artifacts presented as results of the requested generation may have inappropriately appropriated the artistic styles of human professionals. However, we cannot consider it plagiarism, as they do not reproduce an artifact exactly as it originated. In our context, the development of educational games is carried out by students without expertise in design or illustration, voluntarily (without funding). Thus, the use of diffusion models integrated with LLMs is limited to supporting the artistic process and reducing technical limitations during the prototyping phase carried out by computer science students.

As future work, we seek to: (i) expand the study, involving the perspective of illustrators, especially game illustrators; (ii) include more models in the evaluation, thoroughly analyzing their particularities to ensure accurate generation; (iii) generate four or more sprites of a character to simulate animations and investigate the models' ability to maintain visual consistency; (iv) generate sprites from a game still in the idealization phase, to avoid biases; and (v) add more quantitative questions to the questionnaire to aid in a more detailed analysis of the sprites.

References

Asadzadeh, A.; Shahrokhi, H.; Shalchi, B.; Khamnian, Z.; Rezaei-Hachesu, P. Serious educational games for children: A comprehensive framework. **Heliyon**, Elsevier, v. 10, n. 6, 2024.

Battistella, P.; Wangenheim, C. von; Fernandes, J. M. Como jogos educacionais são desenvolvidos? uma revisão sistemática da literatura. In: SBC. **Workshop sobre Educação em Computação (WEI)**. [S.l.], 2014. p. 159–168.

Benjdira, B.; Ali, A. M.; Boulila, W.; Koubaa, A. Integrating large language models and diffusion models in generative ai tasks: Progress, challenges, and future directions. **Authorea Preprints**, Authorea, 2025.

Boller, S.; Kapp, K. **Jogar para aprender: tudo o que você precisa saber sobre o design de jogos de aprendizagem eficazes**. [S.l.]: DVS Editora, 2018.

Bopp, J. A.; Müller, L. J.; Aeschbach, L. F.; Opwis, K.; Mekler, E. D. Exploring emotional attachment to game characters. In: **Proceedings of the Annual Symposium on Computer-Human Interaction in Play**. [S.l.: s.n.], 2019. p. 313–324.

Carvalho, B. R. M. d. **Design de personagens em jogos digitais: uma análise sobre a relação entre narrativa, visualidade e character design aplicada ao jogo League of Legends**. 2024. Trabalho acadêmico. Formato de citação gerado a partir de base bibliográfica.

Chen, H. *et al.* Comprehensive exploration of diffusion models in image generation: a survey. **Artificial Intelligence Review**, Springer, v. 58, n. 4, p. 99, 2025.
Creswell, J. W.; Poth, C. N. **Qualitative inquiry and research design: Choosing among five approaches**. [S.l.]: Sage publications, 2016.
Das, A.; Wu, Z.; Skrjanec, I.; Feit, A. M. Shifting focus with hceye: Exploring the dynamics of visual highlighting and cognitive load on user attention and saliency prediction. **Proceedings of the ACM on Human-Computer Interaction**, ACM New York, NY, USA, v. 8, n. ETRA, p. 1–18, 2024.
Gallotta, R.; Liapis, A.; Yannakakis, G. N. The importance of context in image generation: A case study for video game sprites. In: Springer. **International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)**. [S.l.], 2025. p. 66–81.
Ghosh, A.; Acharya, A.; Saha, S.; Jain, V.; Chadha, A. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. **arXiv preprint arXiv:2404.07214**, 2024.
Grassi, N. B. Aprendizado tangencial e gameflow nos jogos digitais: estratégias para o desenvolvimento de jogos educacionais engajadores. Universidade Estadual Paulista (Unesp), 2021.
Ho, J. C. F.; Ng, R. Perspective-taking of non-player characters in prosocial virtual reality games: Effects on closeness, empathy, and game immersion. **Behaviour & Information Technology**, Taylor & Francis, v. 41, n. 6, p. 1185–1198, 2022.
Hsieh, C.-A.; Zhang, J.; Yan, A. Sprite sheet diffusion: Generate game character for animation. **arXiv preprint arXiv:2412.03685**, 2024.
Jovanovic, M.; Campbell, M. Generative artificial intelligence: Trends and prospects. **Computer**, IEEE Computer Society, v. 55, n. 10, p. 107–112, 2022.
Klopfer, E.; Haas, J.; Osterweil, S.; Rosenheck, L. **Resonant games: Design principles for learning games that connect hearts, minds, and the everyday**. [S.l.]: MIT Press, 2025.
Macena, J.; Melo, G.; Lais, R.; Pires, F.; Pessoa, M. Gramágica: um jogo educativo para praticar classificaç ao silábica através do pensamento computacional. In: **Anais do XVIII Brazilian Symposium on Computer Games and Digital Entertainment**. [S.l.: s.n.], 2019. p. 977–984.
Mayer, R. **Multiple learning**. [S.l.]: New York: Cambridge University Press, 2009.
Merino, T.; Negri, R.; Rajesh, D.; Charity, M.; Togelius, J. The five-dollar model: generating game maps and sprites from sentence embeddings. In: **Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment**. [S.l.: s.n.], 2023. v. 19, n. 1, p. 107–115.
Plass, J. L.; Homer, B. D.; Kinzer, C. K. Foundations of game-based learning. **Educational psychologist**, Taylor & Francis, v. 50, n. 4, p. 258–283, 2015.
Rativa, A. S.; Postma, M.; Zaanen, M. V. The influence of game character appearance on empathy and immersion: Virtual non-robotic versus robotic animals. **Simulation & Gaming**, SAGE Publications, v. 51, n. 5, p. 685–711, 2020.
Sengar, S. S.; Hasan, A. B.; Kumar, S.; Carroll, F. Generative artificial intelligence: a systematic review and applications. **Multimedia Tools and Applications**, Springer, p. 1–40, 2024.
Wu, Z.; Chen, Z.; Zhu, D.; Mousas, C.; Kao, D. A systematic review of generative ai on game character creation: Applications, challenges, and future trends. **IEEE Transactions on Games**, IEEE, 2025.