PicLingo: A GenAI-Based System for Language-Disabled Children

Razan Alatawi[1], Shahad Alamri[2], Renad Almaghthawi[3],
Shada Alofi[4], Ghada Alharbi[5]*, Rehab Albeladi[6]
Department of Computer Science, Taibah University, Medina, KSA[1,2,3,4,6]
Department of AI and Data Science, Taibah University, Medina, KSA[5]

***Abstract*—Children with language disorders often face challenges in understanding sentences and communicating effectively with others. While previous studies have utilized static digital games and automated feedback to support vocabulary and spelling, there remains a significant gap in leveraging generative models to provide dynamic, personalized visual reinforcement for verbal tasks. This study presents a new approach to support language development through PicLingo, a GenAI-powered system developed to assist both children and their mentors. A comparative experimental methodology is used to evaluate multiple generative models using MS COCO samples and standard metrics, including Inception Score (IS), Fréchet Inception Distance (FID), and human evaluation. PicLingo's primary feature is a text-to-image generation (TTI) task that generates illustrative images from textual descriptions. Additionally, the system includes an interactive game that uses speech recognition technology to encourage active verbal participation. This approach aims to enhance language development and overall communication skills. The experimental results demonstrate that the proposed Stable Diffusion-based architecture significantly outperforms baseline models in generating high-quality, semantically accurate images, suggesting PicLingo as a promising, interactive tool for enhancing verbal communication and tracking linguistic progress in children with language disorders**.

***Keywords*—PicLingo; Generative AI; text-to-image generation; speech recognition**

I. Introduction

Generative Artificial Intelligence (GenAI) is opening exciting possibilities in educational technology, offering innovative approaches that have the potential to replace traditional teaching methods. These advancements present unique opportunities to support children with language disorders, who often encounter significant challenges in their learning process. Language disabilities include a variety of conditions, such as language processing difficulties and communication disorders, which can make it difficult to understand vocabulary, sentence structures, and complex phrases, all of which impact a child's ability to express thoughts, interpret information, and communicate effectively.

Visual and game-based approaches have demonstrated considerable efficacy of supporting learning for children with language disorders. The use of images enhances children's comprehension of concepts, whereas games offer interactive and engaging educational experiences. Multiple studies indicate that the utilization of images and games enhances academic performance, social skills, and self-esteem in children with language disorders [1]. The use of games to learn vocabulary words significantly increased scores on vocabulary tests by children with dyslexia, a learning disorder that impacts how people read and interpret words, letters, and symbols [2]. Another study showed that children with autism who learned social skills through images demonstrated notable improvements in their social interactions [3]. These findings support the important role of visual tools in learning, particularly given that research shows visual learning relies on images by up to 90% [4].

This study addresses the limitations of existing static and non-adaptive tools for children with language disorders by proposing an interactive and personalized system. The objective is to bridge the gap between visual learning and verbal communication using generative AI. To achieve this, we conduct a comparative study of generative models and develop PicLingo, a system that integrates text-to-image generation and speech recognition to support language development in an engaging and adaptive environment.

II. Related Work

AI in education has gained popularity over the past few decades and has transformed educational practices. The AI market in the educational sector has grown, and these systems have replaced some roles of instructors [5]. In [6], the authors tested the use of natural language processing (NLP) techniques and automated scoring on the performance of students' scientific writing in a certain subject. Two types of feedback were tested, general and specific feedback. Although the study is limited to the subject of climate change, it has shown how integrating these systems can improve the students' writing practices.

In [21], the authors explored the role of reflection within game-based learning environments (GBLEs) using Crystal Island, an educational game focused on microbiology. The study involved middle school students who were required to diagnose a disease outbreak by engaging in scientific reasoning actions such as information gathering, hypothesis formation, and testing. The results revealed that reflection time increased throughout gameplay, and its timing—particularly when prompts were introduced—significantly influenced the depth of reflection. Furthermore, students' engagement in scientific reasoning and their learning gains moderated this relationship. The findings emphasize the importance of integrating adaptive reflection prompts into GBLEs to dynamically support and foster self-regulated learning and reflective thinking.

The use of such tools for children with learning disabil-

*Corresponding author.

ities can potentially enhance their learning process through educational games, helping to bridge communication gaps in an engaging and supportive way. A study on the effect of intelligent games on developing skills for children with learning disabilities was conducted by Flogie et al. [7]. They analyzed the needs, developed learning content, and created 10 intelligent serious games, then evaluated the proposed methods in a real learning environment. The results showed progress in the communication skills and creativity of participating children.

The study by Girhe et al. [22] introduced Sahayak, an AI-enabled mobile assistive application designed to enhance reading and comprehension skills for children with intellectual disabilities. The system integrates artificial intelligence and augmented reality (AR) to provide an interactive and personalized learning environment. Using AR, the application overlays engaging educational elements onto real-world settings through a smartphone or tablet, allowing children to interact with content in an immersive way. The AI component assesses each child's strengths and weaknesses and dynamically adapts the learning modules to target areas requiring improvement. This individualized approach enables children to progress at their own pace, promoting confidence and sustained engagement in the learning process. The study highlights how the combination of AI and AR technologies can make education more accessible and enjoyable for children with intellectual challenges.

In [8], the authors developed a game based on machine learning (ML) techniques to help children with autism spectrum disorder recognize emotions. The game begins by displaying a silent animation or playing an audio clip that represents an emotion. In the next phase, the child is prompted to express the displayed emotion using their voice. The input audio is then classified using a trained random forest classifier to assess the child's response and provide feedback, achieving a 72% accuracy in emotion recognition.

Emerging technologies such as TTI and speech recognition can significantly contribute to the development of such applications, making them more engaging and interactive. A study proposed by [9] assesses the impact of using speech recognition on writing skills for children with special educational needs. The handwriting was assessed before and after the use of the tool. The results showed improvements not only in the quality and quantity of written text but also in the self-esteem of the child.

According to a recent study, speech-to-text intervention enhances writing production for students with intellectual disabilities [10]. In the intervention session, the teacher presented an image for the child to describe and catch their response with speech recognition technology. The teacher will then read the text to the child to make any corrections. The results demonstrated a strong effect on the child's writing productivity.

Another study discussed using TTI generation to enrich learning material for children [11]. Educators reported increased confidence in performing the tasks related to designing image generation applications for K-12 students, such as generating images for storytelling to support the learning materials and meet the diverse needs of their students.

While existing studies show how AI-driven educational tools can help children with disabilities, despite these advancements, existing systems either lack dynamic content generation or do not integrate both visual and verbal learning in a unified framework. This highlights the need for a system like PicLingo, which combines text-to-image generation and speech recognition to provide adaptive, interactive learning support.

III. Methodology

PicLingo is a website that utilizes TTI generation and speech recognition technologies to provide a serious game for children with language disorders. "Serious games" refers to games designed for educational and cognitive improvement purposes [13]. In this case, PicLingo features a "show-and-tell" game, encouraging children to describe an image verbally to enhance their communication and language skills. This section discusses the detailed implementation of the system.

A. Overall System Architecture

The overall system components and data flow are illustrated in Fig. 1, which shows the interaction between the back-end and front-end of the system. The back-end manages core functionalities and data storage, while the system provides two interfaces: a game-based, engaging interface for children, and another for mentors to manage game content. A detailed description of the back-end and front-end components will be provided.

The front-end was built using a web-based interface to ensure accessibility across devices without requiring installation, which is particularly important for educational settings. The back-end was designed as a separate service to allow independent scaling of the AI components, ensuring that computationally intensive tasks such as image generation do not affect the responsiveness of the user interface.

1) Front-end: The front-end consists of two main workflows, one for mentors and another for children. Mentors are responsible for creating and managing game content to ensure safety. This process begins by entering a prompt, which is validated and then used to generate an image. The generated image is stored in the database for later use in the game. On the other hand, the children's interface displays the "Show-and-Tell" game, where children can select a category from a list (e.g., animals, food, nature, sports, body parts, jobs). After selecting a category, the child views images within it and describes a chosen image verbally using the microphone. This verbal input is then processed by the backend for evaluation.

2) Back-end: The core functions of the system take place in the back-end, including image generation, speech recognition, and similarity computation. Generated images are stored with their captions in a database for future access.

Score computation involves processing the child's spoken input through speech recognition, transcribing it into text, and applying word embeddings to analyze similarity. The system then calculates a score by comparing the transcribed speech with the image caption and returns feedback.

This system architecture is designed to ensure seamless interaction between the front-end and back-end, enhancing the educational impact of the PicLingo system through various

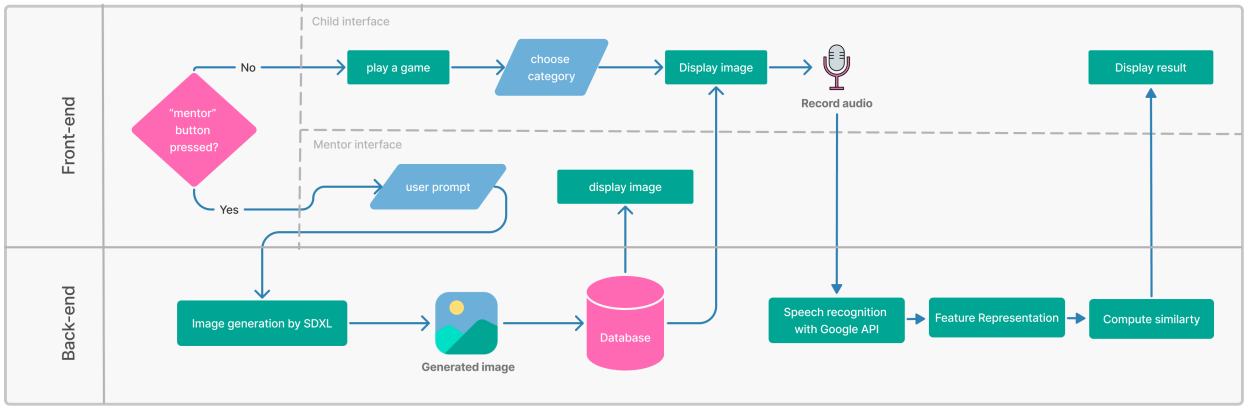

Fig. 1. Flowchart illustrating the system components, data flow, and interactions between the back-end and front-end.

AI-driven features. A discussion of these back-end functions, along with expanded experiments on image generation models, will follow.

B. Text-To-Image Generation and Baseline Models

TTI generation plays a significant role in the PicLingo system's proper functioning to produce visual content based on textual descriptions. To determine which image generation model is most appropriate for PicLingo and better understand their architecture, a comparative study was conducted with baseline and proposed models.

1) Dataset: The two models will be evaluated and compared using a subset of the Microsoft Common Objects in Context (MS COCO) dataset [14]. MS COCO is a widely used large-scale dataset in the field of computer vision, particularly useful for tasks such as object recognition, image segmentation, and image captioning. A notable feature of MS COCO is its image captioning annotations, where each image is paired with multiple human-generated natural language descriptions. The dataset includes a total of 2.5 million labeled objects across 328,000 images. For this experiment, a random subset of 64 samples from this dataset was selected for inference to evaluate the models. This subset size is considered sufficient for testing and model evaluation. While a larger dataset would provide more statistically robust results, the 64-sample subset was considered adequate for the comparative evaluation phase of this study, given that the primary objective was to assess relative model performance rather than absolute generalization capability. We acknowledge this as a limitation and plan to expand the evaluation dataset in future work. Each sample includes one image and one caption. The average number of words per caption is calculated to be 10.48, which reflects the detailed and complex nature of the descriptions provided to the models.

2) Baseline models: We chose the Generative Adversarial Network (GAN) as our baseline model [15], as it is one of the most popular and established models in image generation, making it a strong starting point to compare other models against it. The basic architecture of a GAN model consists of two key components: a generator, which aims to create realistic images, and a discriminator, which works to distinguish between real and generated samples. Our experiments involved the use of several versions of GAN, including BigGAN+CLIP, StyleGAN+CLIP, and VQGAN+CLIP. A comparison of sample images generated by these models is shown in Fig. 2.

The main differences between the models considered lie in their underlying architectures and methodologies. BigGAN+CLIP utilizes a large-scale class-conditioned GAN architecture. StyleGAN+CLIP emphasizes the manipulation of latent space to control specific visual attributes of generated images, resulting in highly customizable outputs [16]. VQGAN+CLIP combines the VQGAN architecture, which incorporates transformers with CLIP, offering a unique approach that excels in generating high-quality images while leveraging semantic guidance from natural language descriptions [17]. These distinctions in architecture contribute to the varying strengths and capabilities between models.

After experimenting with these models, we chose VQGAN+CLIP as the main baseline model due to its promising results and the ability to generate diverse and realistic images, as shown in Fig. 2. The settings used for the experiments were 500 iterations, 256x256 image size, and a version trained on the ImageNet16384 dataset, balancing computational efficiency with image quality. The inference time was averaged at approximately 2 minutes and 30 seconds per image, ensuring practical feasibility within our resources. Table I shows the hyperparameter values used in the experiment.

C. Proposed Model

The proposed model is based on the *Stable Diffusion* architecture, a latent diffusion model capable of generating high-quality, photo-realistic images from textual descriptions [23]. This model operates within a compressed latent space, allowing it to generate images efficiently while maintaining fine detail and semantic consistency. Fig. 3 provides an overview of the main components of the model.

The model employs an 860M-parameter U-Net as its denoising backbone in conjunction with a **CLIP ViT-L/14 text encoder [24]. It was fine-tuned from Stable Diffusion v1-2** and trained for approximately 595,000 iterations at a resolution of 512×512 pixels on the LAION-Aesthetics v2.5+ dataset [25]. During training, a 10% reduction in text

Fig. 2. Samples of images generated using different versions of GANs as baseline models. Bold indicates best results.

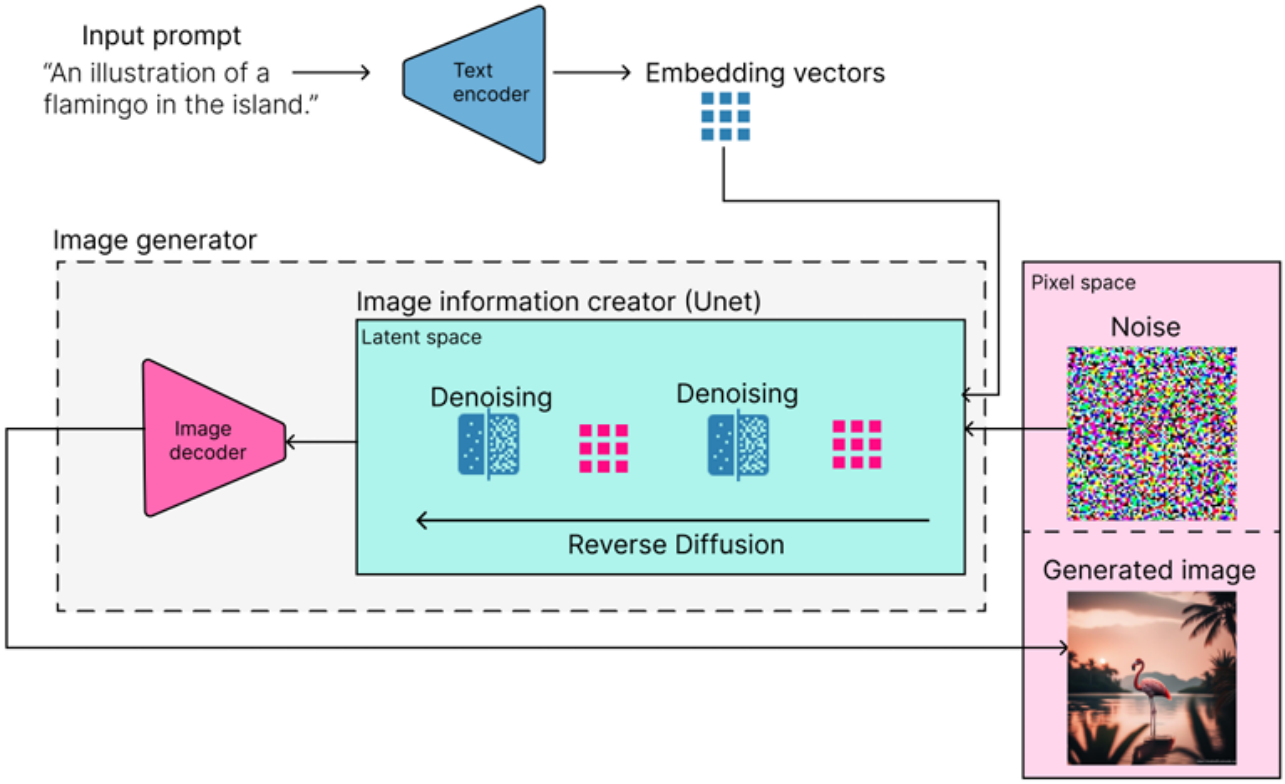

Fig. 3. A flowchart diagram showing the main components of the proposed model.

conditioning was incorporated to improve classifier-free guide sampling [26]. This configuration provides an optimal balance between high performance and computational efficiency. Table II summarizes the hyperparameter settings used in the

TABLE I. HYPERPARAMETER SETTINGS USED IN THE BASELINE MODEL EXPERIMENT.

Hyperparameter	Value
Max iterations	500
Seed	42
Learning rate	0.1
Image resolution	256x256
Optimizer	Adam

TABLE II. HYPERPARAMETER SETTINGS USED IN THE PROPOSED MODEL EXPERIMENT.

Hyperparameter	Value
Number of Inference steps	50
Guidance scale	7.5
Learning rate	1×10^{-4}
Image resolution	512x512
Optimizer	AdamW

experiment.

The selection of Stable Diffusion v1-5 specifically was motivated by three considerations. First, it produces images at a 512×512 resolution, which provides sufficient visual clarity for children to identify and describe objects without overwhelming detail. Second, the guidance scale of 7.5 was chosen after preliminary testing, as lower values produced images that were too abstract for young users, while higher values led to oversaturated outputs that appeared unnatural. Third, the LAION-Aesthetics training data prioritizes visually appealing and coherent images, which aligns with the requirement for clear, recognizable visuals in an educational context for children.

Stable diffusion consists mainly of two main components: the **text encoder**, which processes the input text into a meaningful numerical representation, and the **image generator**, which denoises a latent vector to produce a final image. The denoised latent representation is then passed through an **image decoder** to reconstruct the final high-resolution output.

1) Text encoder: The text encoder is responsible for transforming the input prompt into an embedding that captures its semantic meaning. In this work, the **CLIP (Contrastive Language–Image Pretraining)** model is used as the text encoder. CLIP, developed by OpenAI, was trained on millions of image–text pairs to learn a joint embedding space where both textual and visual concepts are aligned.

The input text is first tokenized and embedded through CLIP's **ViT-L/14 (Vision Transformer–Large, Patch 14)** architecture, which models contextual relationships between words using a transformer-based approach. Each token embedding captures the word's meaning relative to its context, allowing semantically similar concepts to occupy nearby regions in the latent space.

By using CLIP as the text encoder, the diffusion model effectively captures the intent and nuances of the prompt, enabling precise conditioning during the image generation process. This integration of language and vision features allows for more coherent and semantically aligned image synthesis compared to earlier word embedding techniques such as Word2Vec [27], GloVe [28], or BERT [29].

2) Image generator and decoder: The image generator operates within the latent space of the diffusion model. It begins with a random noise vector and iteratively denoises it under the guidance of the text embedding produced by the CLIP encoder. At each diffusion step, the neural network predicts and removes a portion of the noise, gradually refining the latent representation into a semantically coherent visual form aligned with the input text.

An 860M-parameter U-Net serves as the backbone of the denoising network. The U-Net's encoder–decoder structure and skip connections enable it to capture fine-grained details as well as broader spatial structures, making it particularly suitable for diffusion-based image synthesis.

Once the denoising process is complete, the latent vector is passed through a **Variational Autoencoder (VAE) decoder** [28], which reconstructs the latent representation into the pixel space, producing the final image. The VAE decoder ensures that the generated image maintains both high visual fidelity and semantic consistency with the textual prompt.

D. Speech Recognition

Speech recognition technology converts spoken language into text. In the PicLingo system, we use the Google Cloud Speech-to-Text API, renowned for its robust performance as a cloud-based speech recognition tool. The Google Cloud Speech-to-Text API was selected for several reasons specific to this application. First, it supports real-time transcription with low latency, which is essential for maintaining a child's engagement during interactive tasks. Second, it demonstrates robust performance with children's speech patterns, which tend to differ significantly from adult speech in terms of pronunciation and pacing. Third, its cloud-based architecture eliminates the need for local processing resources, making the system more accessible in low-resource educational environments.

E. Cosine Similarity

A text similarity metric calculates the degradation and contextual relevance. In the interactive game, cosine similarity is applied to measure the similarity between two texts, evaluating their semantic meaning, structural competence between the child's description, generated by speech recognition, and the stored image caption.

To implement this metric, the two texts must initially be represented as vectors [12], [30], [31], after which this metric determines the cosine of the angle formed between two vectors, thereby producing a similarity index that ranges from -1 to 1, where a value of 1 indicates complete similarity, a value of 0 indicates the absence of similarity, and a value of -1 indicates complete dissimilarity [see Eq. (1)].

$$\text{Cosine Similarity}(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

Cosine similarity was selected over alternative metrics such as Euclidean distance or Jaccard similarity because it measures the angle between two vectors rather than their magnitude, making it more robust to differences in description length. This is particularly appropriate in this context, as children

with language disorders may produce shorter or less complete descriptions than expected, and penalizing response length would introduce unfair scoring.

IV. Evaluation and Results

This section presents a comparative analysis between the baseline VQGAN model and the proposed Stable-diffusion-v1-5 model, evaluated using the mentioned dataset. To evaluate the performance of both models, three primary metrics were used: Inception Score (IS), Fréchet Inception Distance (FID), and Fleiss' Kappa for human evaluation.

A. Inception Score (IS)

The Inception Score is used to assess both the quality and diversity of generated images [18]. It measures the distance between the distributions of features extracted from these images using a pre-trained image classification model, where lower IS scores indicate better alignment of the generated images with real-world images in the feature space. Before applying IS calculation, images should be preprocessed by converting them to a common RGB format, resizing them to 299x299 pixels to match the expected size for the Inception model, and normalizing pixel values for consistent evaluation conditions. The IS can be calculated using the following Eq. (2):

$$\text{IS}(G) = \exp\left(\mathbb{E}_{x\sim p_g}\left[D_{\text{KL}}(p(y|x)\|p(y))\right]\right), \tag{2}$$

where,

G : generative model being evaluated,
p_g : distribution of generated images,
x : generated images,
y : class labels predicted by an Inception model,
$p(y|x)$: conditional class distribution given the generated images,
$p(y)$: marginal class distribution (often assumed to be uniform),
D_{KL} : Kullback-Leibler divergence.

B. Fréchet Inception Distance (FID)

The FID metric measures the similarity between the feature distributions of real and generated images, with lower scores indicating a closer alignment between these distributions [19]. It takes both the mean and covariance of the distributions into account, making it a robust metric for generative models. This allows FID to capture even small differences between images. FID is calculated using the following Eq. (3):

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{3}$$

where,

μ_r : mean feature vector of real images,
μ_g : mean feature vector of generated images,
Σ_r : covariance matrix of real image feature vectors,
Σ_g : covariance matrix of generated image feature vectors,
$\|\cdot\|$: Euclidean distance,
Tr : trace of a matrix.

C. Fleiss' Kappa for Human Evaluation

To complement the quantitative metrics, a human evaluation was conducted using Fleiss' Kappa, which measures the level of agreement among multiple raters beyond chance. Three raters with backgrounds in computer science were tasked with assessing the relevance of generated images to their respective text prompts, assigning a score of 0 for no match and 1 for a match.

Fleiss' Kappa is calculated based on the observed agreement among raters (Po) and the expected agreement by chance (Pe) using the Eq. (4) [20]:

$$\kappa_{\text{score}} = \frac{P_o - P_e}{1 - P_e} \times 100, \tag{4}$$

Interpretation of Fleiss' Kappa scores indicates the level of inter-rater reliability, with higher scores representing stronger agreement.

Using these three metrics should provide a solid foundation to evaluate the models both qualitatively and quantitatively.

It is important to note that the evaluation was conducted on a relatively small subset of 64 samples due to computational constraints. While this size is sufficient for preliminary comparison, it may limit the statistical strength of the results. Future work will include larger datasets and more comprehensive statistical analysis, including variance reporting, to improve the robustness of the evaluation.

D. Results

The results listed in Table III demonstrated that the proposed model outperformed the baseline model in both IS and FID scores, reflecting the high quality of the generated images. Specifically, the proposed model achieved an IS of 4.26 and an FID of 1.83, while the baseline scored an IS of 4.06 and an FID of 6.44. To support these quantitative results, a human evaluation was conducted as detailed in Section IV-C. The proposed model achieved a kappa of 67%, indicating good agreement on the alignment between generated images and the prompt, whereas the baseline model scored a kappa of 17%, reflecting poor inter-rater agreement. To provide further insights into qualitative outcomes, Fig. 4 presents samples of images generated by both models. These results reveal that the images produced by the proposed model are of higher quality and are more realistic compared to those generated by the baseline. This visual comparison underscores the superior performance of the proposed model in generating visually appealing images, making it a more suitable choice for integration into the system.

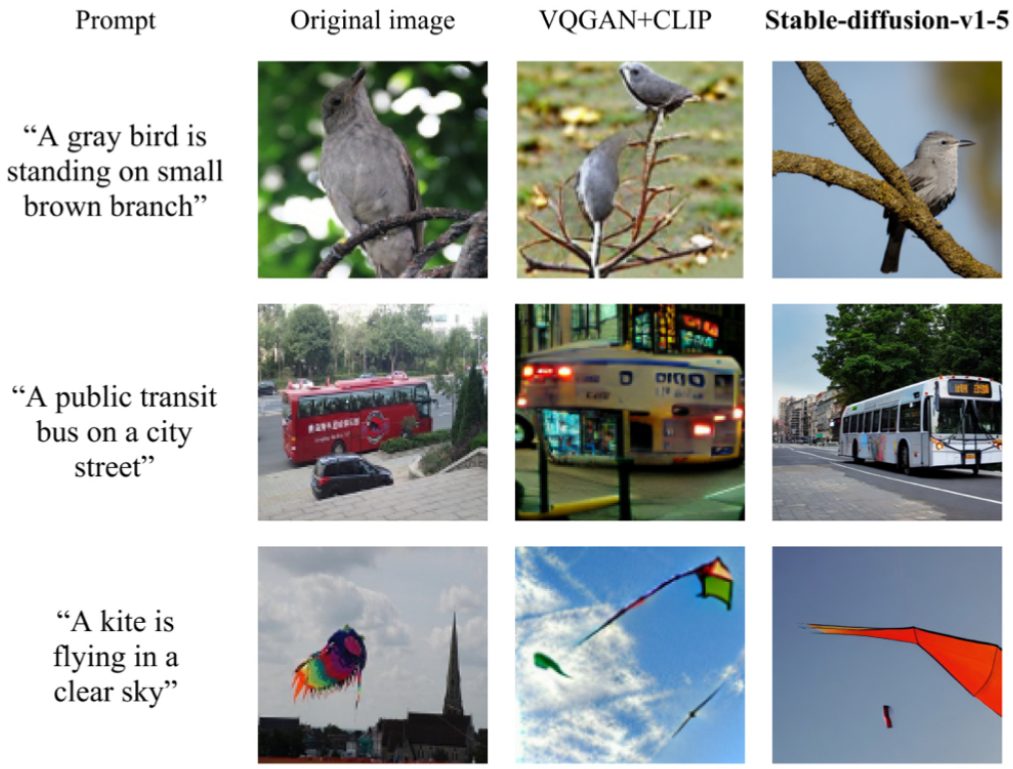

Fig. 4. Samples of images generated using the baseline model versus the proposed model.

TABLE III. IS AND FID SCORES FOR VQGAN AND STABLE-DIFFUSION-V1-5.

Model	IS ↑	FID ↓
Stable-diffusion-v1-5	**4.26**	**1.83**
VQGAN+CLIP	4.06	6.44

V. DISCUSSION

The improvement in image quality and semantic alignment is particularly important for the target application. High-quality and contextually accurate images help children better understand visual concepts and support their ability to describe them verbally, which directly contributes to improving language development and engagement during the learning process.

The results demonstrated that the proposed Stable Diffusion model consistently outperformed the baseline in both quantitative and qualitative evaluations. One key observation is that diffusion-based models better capture semantic alignment between text and image, which is essential for educational applications like PicLingo. This improvement in semantic alignment enables the generation of clearer and more contextually accurate images, which can help children better understand visual concepts and support their ability to describe them verbally. Additionally, the integration of speech recognition with visual feedback created a multimodal learning experience that may enhance engagement and retention in children with language disorders. However, the improvement in performance comes at the cost of higher computational requirements.

From a practical perspective, the system shows promising potential for real-world educational use, particularly in personalized learning environments. However, further validation through real-world user studies is required. Future improvements could focus on optimizing performance and extending support to multilingual settings.

VI. CONCLUSION

This study highlights the potential of Generative Artificial Intelligence (GenAI) in transforming the learning experience of children with language disorders. We, therefore, present a novel pedagogical approach, PicLingo, powered by GenAI to provide a personalized learning environment for both the mentor and the child. Since image generation is one of the important features of this system, we conducted a comparative analysis of prominent generative models in the field to find the most effective model. Our analysis shows that the proposed Stable-diffusion-v1-5 model has superior performance over the baseline model, VQGAN+CLIP, with an IS of 4.26 and an FID of 1.83. Fleiss' Kappa results were 67%. It indicates that The proposed model can generate high-quality and diverse images containing semantic information according to the textual description provided by the mentor. Furthermore, PicLingo includes an interactive game through speech recognition technology that engages children by allowing them to determine whether their descriptions are accurate, making it an effective tool for improving communication skills and tracking a child's language development. Our findings indicate that PicLingo shows promising potential as an assistive educational tool;

however, further validation through real-world user studies is required to confirm its effectiveness in supporting children with language disorders. This highlights the potential of GenAI to enhance learning and contribute to the development of more adaptive educational tools.

A. Limitations and Future Directions

While the results are promising, this study is limited to a controlled environment and a relatively small dataset. Future work will focus on expanding the dataset to include a larger and more diverse set of samples (e.g., over 1,000 images) to improve evaluation reliability and generalizability. Additionally, user studies will be conducted with children and educators to assess the system's impact on language development using pre- and post-assessment methods.

Further improvements will include supporting multilingual capabilities (e.g., Arabic and English) and optimizing system performance to reduce response time in interactive environments. Moreover, integrating multimodal feedback—such as emotional tone recognition and adaptive difficulty adjustment—could further enhance the personalized learning experience.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the College of Computer Science and Engineering at Taibah University for their continuous support and encouragement throughout this research.

REFERENCES

[1] R. Zwitserlood, M. t. Harmsel, J. Schulting, K. Wiefferink, and E. Gerrits, "To game or not to game? efficacy of using tablet games in vocabulary intervention for children with DLD," *Applied Sciences*, vol. 12, no. 3, p. 1643, 2022.

[2] L. Rello, C. Bayarri, Y. Otal, and M. Pielot, "A method to improve the spelling of children with dyslexia," Recuperado a partir de http://www.luzrello.com/Piruletras(Dyseggxia)files/assets20142.pdf, 2014.

[3] G. De Leo, C. Gonzales, P. Battagiri, et al., "A smartphone application and a companion website for the improvement of the communication skills of children with autism: Clinical rationale, technical development and preliminary results," *Journal of Medical Systems*, vol. 35, pp. 703–711, 2011.

[4] M. I. B. Magda M. Saleh, "Developing image reading skills to support visual learning for children with learning disabilities," *Technium Social Sciences Journal*, 2020. –

[5] K. Zhang and A. B. Aslan, "AI technologies for education: Recent research & future directions," Computers and Education: Artificial Intelligence, vol. 2, pp. 100025, 2021. Available: https://doi.org/10.1016/j.caeai.2021.100025.

[6] M. Zhu, O. L. Liu, and H. S. Lee, "The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing," Computers & Education, vol. 143, pp. 103668, 2020. Available: https://doi.org/10.1016/j.compedu.2019.103668.

[7] F. A. Flogie, B. Aberšek, M. K. Aberšek, C. Sik Lanyi, and I. Pesek, "Development and evaluation of intelligent serious games for children with learning difficulties: Observational study," JMIR Serious Games, vol. 8, no. 2, pp. e13190, 2020. Available: https://doi.org/10.2196/13190.

[8] A. Rouhi, M. Spitale, F. Catania, G. Cosentino, M. Gelsomini, and F. Garzotto, "Emotify: emotional game for children with autism spectrum disorder based-on machine learning," in *Companion Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Ray, California, 2019, pp. 31–32. Available: https://doi.org/10.1145/3308557.3308688.

[9] M. Kambouri, H. Simon, and G. Brooks, "Using speech-to-text technology to empower young writers with special educational needs," Research in Developmental Disabilities, vol. 135, pp. 104466, 2023. Available: https://doi.org/10.1016/j.ridd.2023.104466.

[10] C. Sand, I. Svensson, S. Nilsson, H. Selenius, and L. Fälth, "Speech-to-text intervention to support text production for students with intellectual disabilities," Disability and Rehabilitation: Assistive Technology, pp. 1–8, 2024. Available: https://doi.org/10.1080/17483107.2024.2381785.

[11] S. Ali, P. Ravi, K. Moore, H. Abelson, and C. Breazeal, "A Picture Is Worth a Thousand Words: Co-designing Text-to-Image Generation Learning Materials for K-12 with Educators," AAAI, vol. 38, no. 21, pp. 23260-23267, Mar. 2024.

[12] A. Huang et al., "Similarity measures for text document clustering," in Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, vol. 4, pp. 9–56, 2008.

[13] G. A. E. Khayat, T. F. Mabrouk, and A. S. Elmaghraby, "Intelligent serious games system for children with learning disabilities," *2012 17th International Conference on Computer Games (CGAMES)*, Louisville, KY, USA, 2012, pp. 30-34, doi: 10.1109/CGames.2012.6314547.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision—ECCV 2014: 13th European Conference*, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740-755, Springer, 2014.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems (NIPS 2014), Montreal, Canada, December 8-13, 2014, vol. 27, pp. 2672-2680, 2014.

[16] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2085-2094.

[17] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "VQGAN-CLIP: Open domain image generation and editing with natural language guidance," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 88-105, Springer.

[18] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–10, 2017.

[20] J. L. Fleiss, "Measuring nominal scale agreement among many raters," Psychological Bulletin, vol. 76, no. 5, pp. 378–382, 1971.

[21] Dever, D.A., Cloude, E.B., Azevedo, R. (2021). Examining Learners' Reflections over Time During Game-Based Learning. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds) Artificial Intelligence in Education. AIED 2021. Lecture Notes in Computer Science(), vol 12749. Springer, Cham. https://doi.org/10.1007/978-3-030-78270-2

[22] Girhe, T., Pandya, E., Sheejamol, P. T., & Chacko, A. M. (2024). Sahayak: Affordable AI enabled Assistive Technology for Intellectually Disabled. 1–5. https://doi.org/10.1109/spices62143.2024.10779644

[23] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01039

[24] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 8748–8763.

[25] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

[26] Ho, J., and Salimans, T. (2022). Classifier-free diffusion guidance. *NeurIPS 2022 Workshop on Deep Generative Models and Downstream Applications*.

[27] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[28] Kingma, D. P., and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

[29] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[30] Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[31] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.