Communications of the ACM

Multimodal LLMs for Augmentative Communication: Generating Story-Based Pictogram Boards from Images of Drawings, Fictional Characters, and Real-World Scenes

Journal:	*Communications of the ACM*
Manuscript ID	CACM-25-12-5678
Manuscript Type:	Research and Advances
Date Submitted by the Author:	18-Dec-2025
Complete List of Authors:	Lessa Junior, Andre Luis; Instituto de Tecnologia e Liderança, Computer Science Bense Othero, Marilia; Universidade de São Paulo, Faculdade de Medicina da Universidade de São Paulo Mikio Sasaki, Tomaz; Instituto de Tecnologia e Liderança, Computer Science
Keywords:	Artificial Intelligence, Augmentative and Alternative Communication (AAC), Picture Exchange Communication System (PECS), Multimodal AI, Assistive Technology
Computing Classification Systems:	Human-centered computing, Accessibility systems and tools, Accessibility technologies

SCHOLARONE™
Manuscripts

Multimodal LLMs for Augmentative Communication: Generating Story-Based Pictogram Boards from Images of Drawings, Fictional Characters, and Real-World Scenes

ANDRE LUIS LESSA JUNIOR, Instituto de Tecnologia e Liderana, Brazil
MARILIA BENSE OTHERO, Universidade de So paulo, Brazil
TOMAZ MIKIO SASAKI, Instituto de Tecnologia e Liderana, Brazil

Augmentative and Alternative Communication (AAC) systems support children with complex communication needs, however traditional pictogram boards require extensive manual customization. While recent studies have explored automated vocabulary generation using computer vision and natural language processing, the potential of multimodal Large Language Models (LLMs) for creating child-centered AAC materials remains largely unexplored. This study proposes an LLM-based multimodal approach to automatically generate short educational stories and corresponding pictogram boards from images of characters, drawings, and real-world photographs, and evaluates its potential contribution to communication and therapeutic practices for non-verbal autistic children.

CCS Concepts: • **Human-centered computing** → **Accessibility systems and tools**; **Accessibility technologies**.

Additional Key Words and Phrases: Artificial Intelligence, Augmentative and Alternative Communication (AAC), Picture Exchange Communication System (PECS), Multimodal AI; Assistive Technology

ACM Reference Format:
Andre Luis Lessa Junior, Marilia Bense Othero, and Tomaz Mikio Sasaki. 2018. Multimodal LLMs for Augmentative Communication: Generating Story-Based Pictogram Boards from Images of Drawings, Fictional Characters, and Real-World Scenes. *J. ACM* 37, 4, Article 111 (August 2018), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Augmentative and Alternative Communication (AAC) systems play a crucial role in supporting children with complex communication needs and are widely recognized for enabling participation in educational, social, and daily activities. However, fully harnessing the benefits of AAC technologies often requires substantial effort from families, professionals, and the users themselves. According to Waller et al. [**?**], this burden is frequently intensified by poor usability, high learning demands, limited professional expertise, and difficulties in physical access.

Authors' Contact Information: Andre Luis Lessa Junior, Instituto de Tecnologia e Liderana, Sao Paulo, Brazil, andre.junior@sou.inteli.edu.br; Marilia Bense Othero, Universidade de So paulo, Sao Paulo, Brazil, marilia.othero@usp.br; Tomaz Mikio Sasaki, Instituto de Tecnologia e Liderana, Sao Paulo, Brazil, tmsasaki@prof.inteli.edu.br.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

For Review Only

Within AAC systems, one of the most widely adopted approaches is the Picture Exchange Communication System (PECS), which relies on pictograms to help non-verbal children express actions, needs, and preferences [**?**]. Although effective, PECS demands continuous personalization, and many communicative situations fall outside the childs predictable routine. As a result, caregivers and professionals must invest significant time and effort to create and maintain appropriate communication boards.

As highlighted by Di Paola et al. [**?**], AAC is not a one-size-fits-all solution and must be adapted to each childs interests, developmental profile, and communicative environment. This personalization challenge becomes even more complex when children prefer non photographic content such as drawings, sketches, or cartoon characters materials that traditional AAC tools and computer vision pipelines frequently struggle to interpret.

To address these limitations, this exploratory study investigates whether multimodal Large Language Models (LLMs) can overcome constraints observed in prior computer visionbased approaches to AAC content generation. Studies such as Vargas et al. [**?**] highlight limitations in the semantic interpretation of drawings and fictional characters by traditional vision models. In contrast, this work explores how multimodal LLMs can generate short, child-centered stories and corresponding pictogram boards from images of drawings, fictional characters, and real-world photographs. By integrating visual understanding with language generation, the proposed approach examines the potential of next-generation AI models to support more flexible and engaging AAC materials for non-verbal children.

For Review Only

2 Related Works

Recent studies have explored diverse approaches to the integration of Artificial Intelligence (AI) within Augmentative and Alternative Communication (AAC) systems, particularly focusing on tools designed to enhance communicative accessibility and adaptability for users with communication challenges. The review was developed based on searches conducted through Google Scholar and Scielo.

Di Paola et al.[**?**] introduce AMBRA, an intelligent framework that applies text-to-image and text-to-text generation techniques to automate the customization of communication boards. This approach enhances the adaptability and contextual relevance of content within AAC environments. Camilo et al. [**?**] use logistic regression to predict the need for Picture Exchange Communication System (PECS) implementation. Park et al. [**?**] introduce PICTALKY, converting text or audio into pictograms. Neamtu et al. [**?**] propose LIVOX, a speech-to-text mobile app, which predicts pictograms through contextual AI, improving communication flow.

More recent work by Vargas et al. [**?**] combines computer vision and natural language processing to interpret user uploaded photos, demonstrating encouraging results in image comprehension and text generation for AAC contexts. The study also highlights opportunities for further refinement, such as expanding recognition to include cartoons and stylized drawings, as well as enhancing adaptations for different age groups [**?**]. Inspired by these directions, the present study introduces a multimodal LLM based approach designed to broaden image understanding and provide vocabulary tailored to children through short stories, contextually meaningful sentences.

3 Methodology

The study adopted an exploratory research design aimed at investigating **the extent to which multimodal Large Language Models can generate semantically coherent AAC stories and pictogram boards from drawings and fictional characters in contrast to traditional computer vision approaches**. The proposed system automatically generates short, child-appropriate stories and the corresponding set of pictograms directly from an input image, resulting in a communication board tailored to the visual context presented.

To address this research question, the methodology was conducted in two stages. First, the system was developed and tested using different multimodal LLMs in order to assess their ability to interpret drawings and fictional characters and generate AAC-compatible content. Second, a pilot study was carried out with professionals experienced in AAC, who interacted with the platform and provided feedback through a structured questionnaire combining quantitative ratings and qualitative observations.

The choice of multimodal LLMs is motivated by recent advances in foundation models developed by major AI research groups, which have significantly improved visual language understanding across diverse domains. Traditional computer vision approaches often rely on large, domain specific datasets to generate captions or semantic labels, making them less effective when applied to drawings and fictional characters. In contrast, multimodal LLMs benefit from broad pre-training on heterogeneous data and can generalize to novel visual concepts with minimal or no task-specific fine-tuning. While open-source models may require fine-tuning to adapt to specific AAC contexts, proprietary models can often perform effectively across multiple domains without additional training.

This study does not aim to evaluate long-term therapeutic outcomes. Instead, it focuses on assessing the feasibility of the proposed system from the perspective of professionals experienced in AAC, examining whether the platform is considered appropriate and usable within educational and therapeutic contexts. In addition, the study explores the applicability of multimodal Large Language Models in this setting by evaluating their ability to generate semantically coherent stories and pictogram boards through the developed system. Together, these aspects allow an initial assessment of both the systems practical viability and the suitability of multimodal LLMs for AAC content generation involving drawings and fictional characters.

3.1 Professionals Questionarie

To analyze the perceptions of professionals, the responses to the questionnaire were organized into four evaluative categories: pictogram and story quality, usability and system performance, therapeutic impact and integration and professional applicability. All elements were rated using a four-point Likert scale ranging from 1 (insufficient) to 4 (fully satisfactory). The complete set of questionnaire items is presented below.

Q1 How do you rate the quality of the pictograms presented?
Q2 How do you rate the clarity and adequacy of the generated stories?
Q3 The application interface is easy to understand and use.
Q4 The clarity of captions and pictograms meets the needs of clinical practice.
Q5 The systems response time is adequate for the therapeutic context.
Q6 The tool helped increase the childs interest and attention during the session.
Q7 The resource facilitates the childs communication or expression.

For Review Only

Table 1. Questionnaire categories and corresponding items

Category	Included Questions
Pictogram and Story Quality	Q1, Q2
Usability and System Performance	Q3, Q5
Therapeutic Impact	Q4, Q6, Q7, Q8
Integration and Professional Applicability	Q9, Q10, Q11

Q8 The application reduced signs of frustration or interaction difficulties.

Q9 The resource complements other AAC strategies already used in therapy.

Q10 Overall, I consider the tool useful for therapeutic work.

Q11 I believe this resource can generate long-term benefits for the childs communication.

The correspondence between categories and questionnaire items is summarized in Table 1.

3.2 Multimodal LLM Selection and Prompting

The system relies on a multimodal Large Language Model (LLM) to generate short, context-specific stories from an input image, which are subsequently used to retrieve AAC pictograms. Multimodal LLMs were selected due to their ability to jointly interpret visual and linguistic information, enabling greater semantic flexibility when processing heterogeneous inputs such as real-world photographs, childrens drawings, and fictional characters.

To ensure compatibility with AAC/PECS constraints, a fixed instruction prompt was used to control the structure and linguistic properties of the generated stories. The prompt constrained the model to produce short, developmentally appropriate narratives in Portuguese, using simple vocabulary suitable for pictogram-based communication. This instruction was consistently applied to all images submitted to the system and directly influenced the textual output used for pictogram retrieval.

Two multimodal LLMs were considered during system development: Gemini 2.0 Flash (Google) and the open-source LLaVA-Next [**?**]. Both models were tested to verify their feasibility for integration into the proposed pipeline, focusing on their ability to interpret drawings and fictional characters and to generate AAC-compatible narratives. Based on this preliminary assessment, Gemini 2.0 Flash was selected as the primary model for the system due to its more stable performance under the imposed constraints.

4 Result

4.1 System Overview

The proposed system was designed to be applicable across different usage contexts and accessible from multiple devices. For this reason, it was implemented as a responsive web application, allowing use on desktops, tablets, and mobile devices without platform-specific dependencies.

Considering the exploratory nature of the study and cost constraints commonly associated with pilot applications, the system architecture was designed with operational efficiency in mind. The back-end was deployed on Google Cloud Run, which enables on-demand execution and remains inactive when no requests are being processed, reducing infrastructure costs. The front-end was implemented using React and TypeScript and deployed on Vercel, providing a lightweight and scalable interface for user interaction.

For Review Only

Figure 1 illustrates the overall architecture of the system, including the web interface, the multimodal LLM responsible for story generation, and the pictogram retrieval pipeline.

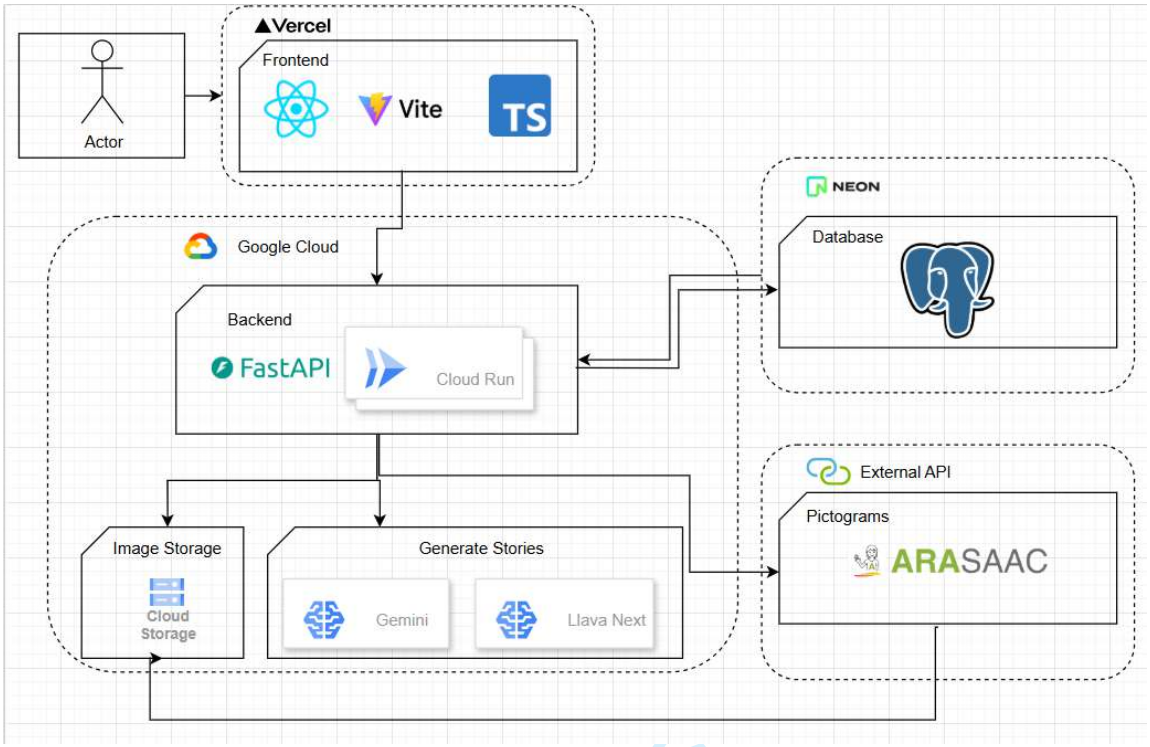

Fig. 1. Overview of the proposed system architecture.

4.2 Multimodal LLM Results

The evaluation of the multimodal LLM component focused on its ability to recognize fictional characters and generate AAC-compatible stories from images. Two models were tested during system development: Gemini 2.0 Flash and the open-source LLaVA-Next.

In the character recognition task, both models correctly identified well-established characters such as Naruto, Sakura, Sasuke, Superman, and Sonic. However, differences emerged when processing more recent or preschool-oriented content. LLaVA-Next frequently failed to recognize characters such as Baby Shark and Luna from the cartoon *Earth to Luna!*, producing incomplete or unrelated descriptions. In contrast, Gemini consistently identified all tested characters across images from different generations of childrens media.

Regarding story generation, Gemini produced short, coherent, and developmentally appropriate narratives that complied with AAC constraints, including sentence length limits and simplified vocabulary suitable for pictogram-based communication. The generated stories remained aligned with the visual content of the images. LLaVA-Next, on the other hand, often violated the imposed constraints by generating overly long descriptions or narratives that lacked semantic alignment with the image.

Table 2. Mean scores per category and professional

Category	Prof. 1	Prof. 2	OT 1	Physio	OT 2
Pictogram and Story Quality	2.5	3.5	2.5	3.0	2.0
Usability and System Performance	3.5	4.0	4.0	3.0	3.5
Therapeutic Impact	3.7	3.8	3.7	3.5	2.5
Integration and Professional Applicability	4.0	4.0	3.7	4.0	3.7

Note: Prof. = Professor; OT = Occupational Therapist; Physio = Physiotherapist.

Based on these results, Gemini 2.0 Flash demonstrated greater robustness and reliability for character interpretation and AAC-oriented story generation within the proposed system.

4.3 Pilot Evaluation

The pilot evaluation was conducted with five professionals experienced in AAC, including teachers and occupational therapists.

Table 2 summarizes the mean scores obtained across the evaluation categories in the professional pilot study. Overall, higher average scores were associated with system usability and professional applicability, while content-related aspects presented comparatively lower averages.

At the item level, the highest scores were observed for questions related to system response time (Q5), overall usefulness for therapeutic work (Q10), and perceived long-term communication benefits (Q11). In contrast, the lowest score was assigned to the item concerning story clarity and adequacy (Q2).

Qualitative feedback indicated that, in many cases, the system generated two stories with different levels of alignment: the first story was generally coherent with the input image and appropriate for children, whereas the second story often resembled a generic childrens narrative with limited or no connection to the visual context.

5 Discussion

5.1 Model Choice and Narrative Accuracy

The results indicate that model choice affected both character recognition and narrative alignment. The proprietary multimodal model Gemini 2.0 Flash demonstrated higher accuracy in identifying childrens fictional characters and generating visually grounded stories. In contrast, the open-source model LLaVA-Next showed limitations when handling more recent characters and occasionally produced narratives weakly connected to the input image.

These findings do not diminish the relevance of open-source models. Instead, they suggest that LLaVA-Next would benefit from additional domain-specific training data focused on contemporary childrens characters, as well as from more structured prompting or narrative templates to improve story consistency.

5.2 Multimodal LLM Advantages Over Computer Vision

The results highlight that multimodal LLMs offer distinct advantages over traditional computer vision models, particularly in recognizing drawings and childrens characters. Unlike computer vision models, which

For Review Only

struggled to identify non-photorealistic images or fictional characters, the multimodal LLMs demonstrated the ability to accurately process both drawings and fictional characters without requiring fine-tuning.

Moreover, the LLMs were able to adapt story generation to be more child-friendly, tailoring narratives to a level suitable for young audiences. This capability stems from the large-scale, diverse datasets used to train LLMs, allowing them to generalize across various visual domains, whereas computer vision models are often limited by domain-specific training data.

5.3 Web-Based Deployment and Professional Adoption

The results indicate that the web-based implementation of the system was well received by professionals, particularly due to its fast response time and ease of access across devices. The platform was perceived as a practical tool that can complement existing AAC approaches without introducing additional technical complexity into professional workflows.

In professional feedback, one participant reported having informally explored the system in a real-world educational context, noting that the models ability to identify elements within an image and generate a corresponding story contributed to increased engagement. This observation highlights the potential of multimodal LLM-based systems to generate context-aware narratives from everyday visual content, reinforcing their applicability as complementary resources in AAC-oriented activities.

5.4 Prompting and Semantic Constraints as Key Limitations

Based on the qualitative analysis of professional feedback and open-ended comments regarding the application, the results indicate that the moderate scores (2.7/4.0) assigned to story and pictogram quality were not primarily related to limitations of the multimodal LLM itself, but rather to design choices in prompting and semantic retrieval. Professionals consistently reported that the first generated story was coherent and well aligned with the input image, whereas the second story often lacked visual grounding. This pattern suggests that the prompt structure did not sufficiently constrain the model to remain anchored to the visual content, pointing to the need for prompt refinement rather than model replacement.

Similarly, qualitative comments highlighted inconsistencies in pictogram selection associated with semantic ambiguity at the word level. In some cases, a single word corresponded to multiple meanings depending on context, leading to the retrieval of inappropriate pictograms. These observations underscore the necessity of incorporating an additional semantic validation step to ensure contextual alignment between generated words and pictograms. Together, these findings suggest that targeted improvements in prompt design and semantic filtering could substantially enhance content quality without changes to the underlying LLM.

6 Conclusion

This exploratory study investigated the extent to which multimodal Large Language Models can generate semantically coherent AAC stories and pictogram boards from drawings and fictional characters, in contrast to traditional computer vision approaches. The results demonstrate that multimodal LLMs are capable of interpreting non-photographic visual inputssuch as drawings and cartoon charactersand generating AAC-oriented content.

The findings indicate that this capability is not uniform across models. Proprietary multimodal LLMs, such as Gemini 2.0 Flash, showed higher accuracy in character recognition and greater narrative alignment

For Review Only

without requiring task-specific fine-tuning, likely due to continuous updates and exposure to large-scale, heterogeneous training data. In contrast, open-source models, while still viable, would require additional domain-specific datasetsparticularly involving contemporary fictional charactersand more structured output constraints to reach comparable performance in this context.

The study also reveals that semantic coherence is highly sensitive to system design choices, particularly prompt formulation. Although the first generated story was generally well aligned with the visual input, the inclusion of a second, less constrained story frequently led to semantic drift. This issue had a measurable impact on professional evaluations, resulting in moderate scores for story quality despite otherwise positive system performance. These findings underscore the importance of precise prompting and visual grounding in AAC-oriented applications, where even a single incoherent narrative can reduce perceived reliability.

Overall, this work provides evidence that multimodal LLMs can meaningfully support the generation of AAC stories and pictogram boards from drawings and fictional characters. Based on professional feedback, the proposed system was perceived as promising for both therapeutic and educational interventions, particularly as a complementary tool within existing AAC practices. However, the extent of this contribution depends on the choice of model and on careful system-level design, particularly regarding prompt constraints and semantic filtering. Future work should therefore focus on refining prompting strategies, improving semantic validation, and exploring hybrid solutions that combine the adaptability of multimodal LLMs with domain-specific controls to enhance reliability in AAC applications.

Received 18 December 2025; revised ; accepted

For Review Only