A Review on Multilingual Speech-to-Video Generation for Hearing-Impaired Object Learning

Sandesh PY
Dept. of CS & Engineering
Vidyavardhaka College of Engg.
Mysuru, India
vvce24cse0583@vvce.ac.in

Dr. Hamsaveni M
Dept. of CS & Engineering
Vidyavardhaka College of Engg.
Mysuru, India
hamsaveni.m@vvce.ac.in

Dr. Chethana H T
Associate Professor
Dept. of CS & Engineering
Vidyavardhaka College of Engg.
Mysuru, India
chethanaht@vvce.ac.in

***Abstract*—The rapid advancement of generative artificial intelligence offers a unique and unprecedented opportunity to bridge the communication and learning gap for children with hearing impairments. This paper presents a comprehensive, in-depth review of recent developments in speech-driven visual synthesis, specifically analyzing their potential for targeted educational applications rather than general media consumption. While significant research over the past decade has focused heavily on generating complex fictional narratives, stylistic artistic content, or realistic lip-synchronizing avatars for the entertainment sector, there remains a critical lack of systems purposefully designed to aid in fundamental vocabulary acquisition—specifically, the visual learning of everyday objects. Existing assistive tools often overlook the necessity for simple, direct word-to-visual translation, especially in the context of regional Indian languages, which leaves a vast demographic digitally marginalized. This review thoroughly evaluates the state-of-the-art literature to identify a major technological and pedagogical gap: the absence of accessible, multilingual frameworks that can instantly and accurately convert a spoken or typed word (in languages such as English or Kannada) into a clear, dynamic, and cognitively appropriate video representation of that specific object. Based on this exhaustive analysis, the paper formulates a robust problem statement and conceptualizes a proposed system architecture. This proposed system aims to serve as an interactive cognitive learning aid, allowing special-needs children to visually observe the words they cannot hear. By bypassing traditional static flashcards and utilizing immersive video generation, the framework seeks to significantly reinforce object recognition, accelerate language retention, and provide a more inclusive technological foundation for early childhood special education.**

***Index Terms*—Speech-to-Video, Generative Artificial Intelligence, Special Education, Hearing Impaired, Object Learning, Multilingual Processing, Natural Language Processing.**

I. Introduction

The acquisition of fundamental language skills and vocabulary is a critical milestone in early childhood development. For children with hearing impairments, learning the names, concepts, and physical behaviors of everyday objects presents a profound and multifaceted challenge. Typical childhood learning methodologies rely heavily on continuous auditory reinforcement—for example, a child hears the spoken word ”apple” simultaneously while seeing the physical fruit or a picture of it. This continuous audiovisual loop builds strong neural associations between phonetic sounds and semantic meanings. Without this auditory loop, hearing-impaired children are forced to rely almost entirely on visual and tactile cues to understand their environment.

Historically, special education has relied on traditional, non-digital learning aids such as printed flashcards, static picture books, and manual sign language instruction. While these methods are foundational and highly useful, they possess inherent limitations. Static images fundamentally lack the dynamic nature of the real world. A printed picture of a ”dog” can show its shape and color, but it cannot convey the action of the dog running, barking, or playing—actions that are deeply tied to the cognitive understanding of the noun itself. Furthermore, static mediums often fail to capture and sustain a young child’s attention for prolonged educational periods, leading to slower vocabulary acquisition rates compared to their hearing peers.

Recent exponential strides in artificial intelligence, particularly in the subfields of computer vision and natural language processing, have made it entirely possible to generate highly realistic, high-resolution images and videos directly from simple text prompts. Latent diffusion models and generative adversarial networks have revolutionized media creation. However, the vast majority of these state-of-the-art technologies are architected and optimized for creative professionals, cinematic visual effects, or general entertainment. They are frequently designed to hallucinate elaborate scenes, surreal artistic styles, or complex fictional narratives. Consequently, they rarely address the specific, practical, and highly constrained needs of special education. In a learning environment, the primary goal is semantic clarity and immediate, unambiguous association, rather than artistic complexity or cinematic flair.

Furthermore, the development and training of these massive artificial intelligence models are predominantly English-centric. The underlying datasets used to train text-to-video models consist overwhelmingly of English text-video pairs. This linguistic bias leaves out regional languages like Kannada, creating a severe secondary barrier for local students in regions like Karnataka, India. A child in a semi-urban or rural setting relies on their native tongue for basic cognitive scaffolding. Forcing an English-only technological interface introduces an unnecessary and often overwhelming cognitive load, directly impeding the primary educational goal of object recognition.

The primary objective of this paper is to critically review the current state of speech-to-visual technologies and propose a highly focused, accessible solution tailored specifically for educational accessibility. The goal is to define the architectural requirements for a system that acts as a dynamic visual dictionary. In this proposed framework, a user can seamlessly speak or type the name of an object in their native language and immediately watch a generated, accurate video of that object in action. This approach intentionally shifts the technological focus away from complex, long-form storytelling towards targeted concept visualization, thereby providing a powerful, inclusive new tool for cognitive development.

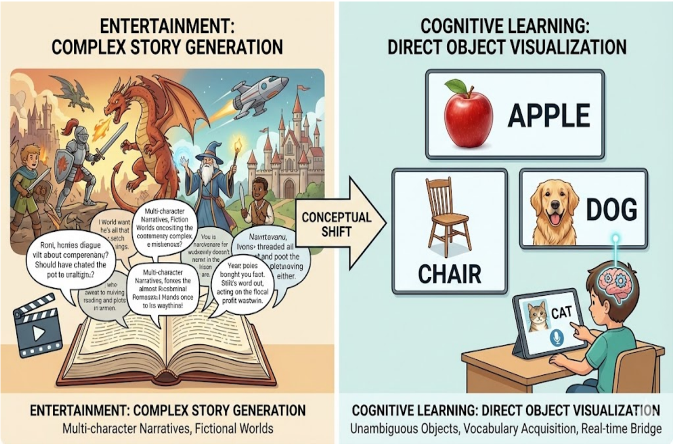
ENTERTAINMENT:
COMPLEX STORY GENERATION
COGNITIVE LEARNING:
DIRECT OBJECT VISUALIZATION
APPLE
CHAIR
DOG
CAT
CONCEPTUAL SHIFT
ENTERTAINMENT: COMPLEX STORY GENERATION
Multi-character Narratives, Fictional Worlds
COGNITIVE LEARNING: DIRECT OBJECT VISUALIZATION
Unambiguous Objects, Vocabulary Acquisition, Real-time Bridge

Fig. 1: Conceptual shift in generative applications: Moving from complex, multi-character story generation designed for entertainment to direct, unambiguous object visualization explicitly designed for cognitive learning. The system acts as a real-time visual bridge for vocabulary acquisition.

II. Literature Review

To fully grasp the technological landscape and understand why a dedicated object-learning framework is strictly necessary, it is essential to examine key research in the intersecting fields of generative media, natural language processing, and digital accessibility. The existing literature can be broadly categorized into distinct domains, each contributing a piece of the puzzle but ultimately falling short of a unified educational solution.

A. Visual Consistency in Generative Media

In the context of visual learning, the generated output must accurately and consistently represent the spoken word without introducing confusing visual artifacts, morphing objects, or irrelevant backgrounds. Patrick Kwon et al. [1] developed a pipeline known as ”DreamingComics,” which focused heavily on maintaining visual consistency of subjects across multiple comic panels. By utilizing modified video diffusion models and strict layout controls, their work successfully ensures that characters retain their specific visual identity in different generated scenes. However, while technologically impressive, the output format of a comic strip remains inherently static. For a hearing-impaired child learning the concept of a ”dog,” a moving video that accurately captures the animal’s physical behavior is exponentially more instructive than a series of static drawings.

Addressing the need for temporal stability in moving media, Yupeng Zhou et al. [3] introduced ”StoryDiffusion.” This methodology implements advanced self-attention mechanisms within the generation pipeline to keep characters and environments consistent over longer periods of time, solving the common generative issue where objects randomly change shape or color from frame to frame. These technologies form a crucial foundation for any visual system; however, the current implementations are overly complex. They need to be fundamentally simplified and optimized to focus on generating single, isolated objects rapidly, rather than rendering complex, slowly evolving character arcs that require massive computational overhead. Furthermore, research by Jezia Zakraoui et al. [4] explored text-to-scene layouts utilizing natural language processing to parse spatial relationships, but their output similarly lacks the dynamic motion necessary for full cognitive grasping of verbs and actions associated with objects.

B. Bridging Text to Educational Visuals

The semantic bridge between user text prompts and the resulting visual output is a critical area of study. Zhen Bin It et al. [2] explored this connection by utilizing large language models to automatically expand simple, vague prompts into highly detailed, descriptive scripts suitable for video generation. In the proposed educational context, this capability is highly relevant for adding necessary visual context to a single isolated word. For example, if a child simply says ”ball,” the raw generative model might produce an unpredictable or abstract image. An intermediary language model can understand the educational intent and expand the prompt to ”a bouncing, brightly colored rubber ball on a grassy field,” thereby guaranteeing a lively, recognizable video. However, existing research in this specific area overwhelmingly focuses on generating creative fiction and automated movie scripts rather than ensuring the factual, clear, and unambiguous representation required for foundational special education. The risk of hallucination—where the artificial intelligence generates physically impossible or semantically incorrect visuals—must be tightly constrained in a learning environment.

C. Speech Processing and Auditory Accessibility

For an assistive tool to be truly accessible to its target demographic, it must possess highly accurate voice input processing capabilities, capable of understanding users with diverse speech patterns. Sreevathsa Sree Charan et al. [9] developed advanced real-time speech-to-text systems, heavily emphasizing the vital importance of low-latency, accurate transcription for deaf individuals. However, their research fundamentally stops at the transcription phase. For a young hearing-impaired child who is still developing reading skills, text alone is often entirely insufficient; they urgently require the visual equivalent of the transcribed text.

Other researchers have attempted to visualize speech directly. Brayan Bernardo et al. [6] and K. L. Liu et al. [7] focused their efforts on utilizing transformer neural networks and emotion mapping algorithms to create virtual talking heads and emotionally expressive facial animations. While these systems represent a massive leap in human-computer interaction and are incredibly impressive technically, they serve a completely different accessibility need. They primarily aid in lip-reading and recognizing human facial emotions rather than teaching the user about the external physical world. A highly realistic, lip-synchronizing talking face cannot teach a child what the physical characteristics of a ”helicopter,” a ”tiger,” or a ”mountain” are.

D. Multimodal Learning Tools and Deployment

The pedagogical implications of these tools are vast. Corina-Marina Mirea et al. [10] discussed the broader systemic impact of integrating generative artificial intelligence into modern educational frameworks. Their research argues convincingly that customized, artificially generated media significantly increases student engagement and retention rates compared to standardized curricula. This directly supports the core premise of this paper: a custom, dynamic video generated instantly for a specific word a child wants to learn is far more memorable than a generic, static textbook image.

Despite this, the deployment of such systems remains a hurdle. Hyungjun Doh et al. [8] investigated multimodal generative artificial intelligence within storytelling but noted that many immersive tools require expensive, specialized hardware like augmented reality glasses or high-end graphics processing units. This hardware dependency makes such solutions completely inaccessible to the average public school classroom, special education center, or low-income household.

The comprehensive review of the current literature indicates a glaring, unaddressed gap. While powerful, isolated tools for speech recognition, language translation, and video generation exist independently, they have not been logically integrated into a simple, lightweight, multilingual educational tool. Current state-of-the-art systems are either far too computationally complex (designed for generating full cinematic movies) or far too narrow in scope (designed for generating only human faces). They entirely miss the necessary, practical middle ground of rapidly generating simple, unambiguous object concepts for daily visual learning.

III. Comparative Analysis

To systematically understand the limitations of the current technological landscape regarding the proposed educational goals, Table I presents a detailed comparative analysis. This analysis evaluates all reviewed papers, highlighting their core underlying technologies, their primary scientific contributions, and critically, why they fall short of solving the specific problem of real-time, multilingual object visualization for hearing-impaired children.

The comparative analysis clearly illustrates a divergence between the direction of current artificial intelligence research and the practical needs of special education. While the underlying mathematical models (such as transformers and diffusion networks) are immensely powerful, their application layers are misaligned with the requirements of accessible learning. The computational overhead required by systems designed to maintain long-term narrative consistency (for example, StoryDiffusion) renders them too slow for a real-time classroom environment where a child expects immediate visual feedback. Conversely, systems optimized for real-time performance (for example, Speech-Driven Talking Heads) are too narrow in their visual scope to act as a comprehensive visual dictionary.

IV. Problem Statement

Synthesizing the extensive literature review and the comparative analysis makes it abundantly clear that while the disparate software components for a speech-to-learn pipeline exist in isolation, they have absolutely not been integrated into a cohesive, lightweight, and purpose-built tool for special education.

The fundamental core problem is that hearing-impaired children currently lack a dynamic, interactive, and independent method to learn new physical vocabulary in their native regional language. To acquire new concepts, they are perpetually forced to rely on scarce human sign language interpreters, static unengaging textbooks, or digital platforms that are strictly limited to the English language. This creates a severe bottleneck in early childhood cognitive development and exacerbates educational inequality.

To directly combat this issue, this paper formally defines the critical need for a **”Multilingual Visual Learning Assistant”**. To be successful, deployable, and impactful, the proposed software system must strictly adhere to the following architectural and functional requirements:

- **Dual-Mode Omnichannel Input:** The system architecture must seamlessly accept both voice input (crucial for children who possess vocal abilities but suffer from auditory processing deficits) and text input (necessary for older children who are actively

TABLE I: In-Depth Analysis of Current Technologies versus Specific Educational Needs

Ref	Paper Title	Core Technology	Primary Contribution	Limitation for Object Learning Application
[1]	DreamingComics	Video Diffusion Models	Achieved highly consistent character visuals across static panels.	Static images are inherently less engaging than dynamic video for learning verbs, actions, and object behaviors.
[2]	Generative Artificial Intelligence Script Writing	Large Language Models	Automates the creative script generation process from vague prompts.	Focuses overwhelmingly on complex fiction and multi-character stories, rather than clear, isolated object definition.
[3]	StoryDiffusion	Self-Attention Mechanisms	Guarantees long-term visual consistency in continuous video generation.	Designed primarily for long cinematic narratives, which requires excessive computational time for instant, single-object visualization.
[4]	Pipeline for Story Visualization	Natural Language Processing	Effectively converts complex text descriptions to spatial scene layouts.	Lacks integrated audio input processing; outputs remain static and lack the dynamic representation needed for engagement.
[5]	Character Animation Survey	Motion Synthesis Algorithms	Provides a comprehensive overview of realistic versus stylized motion generation.	Focuses almost entirely on human bipedal characters, severely neglecting the generation of general inanimate objects or animals.
[6]	Speech-Driven Talking Head	Transformer Neural Networks	Delivers precise, real-time lip-synchronization for digital human avatars.	Only visualizes human faces; cannot be utilized to teach external environmental object vocabulary.
[7]	Speech-driven Cartoon Animation	Emotion Mapping Algorithms	Successfully maps human speech prosody to two-dimensional cartoon facial expressions.	Focuses exclusively on character facial emotions, failing to provide a mechanism to visualize external nouns or objects.
[8]	Multimodal Augmented Reality	Multimodal Generative Artificial Intelligence	Significantly enhances storytelling immersion through environmental integration.	Requires highly specialized, expensive augmented reality hardware, drastically reducing widespread accessibility for average users.
[9]	Real-Time Speech-to-Text	Automatic Speech Recognition	Converts spoken words to highly accurate textual transcripts in real-time.	Provides raw text output only; entirely lacks the generative visual component strictly required for cognitive concept grasping.
[10]	Transforming Learning	Educational Artificial Intelligence Frameworks	Demonstrates quantitatively that artificial intelligence tools increase general student engagement.	Remains a high-level theoretical framework; lacks a specific, deployable software implementation for vocabulary acquisition.

learning to read and write and prefer keyboard interaction).

- **Robust Regional Language Support:** It is absolutely critical to break the English-only barrier. The system must natively support regional languages, specifically Kannada, alongside English. The linguistic backend must ensure that a child speaking the Kannada word ”ಪುಸ್ತಕ (Pustaka)” triggers the exact same internal semantic mapping and visual video output as a child speaking the English word ”Book”.
- **Direct, Unambiguous Object Visualization:** The definition of system success is strictly redefined. Success is not generating a complex, award-winning cinematic story. Success is generating a clear, immediately recognizable, and unambiguous short video of the requested object in a neutral environment. If the user input is ”Elephant”, the required output is a high-fidelity video of an elephant walking or eating, effectively serving as an interactive digital flashcard.
- **Platform Accessibility and Low Latency:** The final system must be deployed as a highly optimized web-based application. It must be exceptionally simple to use and require absolutely no specialized hardware, augmented reality headsets, or local high-end graphics processing units. It must function smoothly on a standard low-cost school computer or a parent’s smartphone, relying on cloud-based processing to deliver the video output rapidly.

V. Proposed Methodology (Conceptual)

To effectively solve the deeply entrenched problems identified in the previous sections, a novel, modular con-

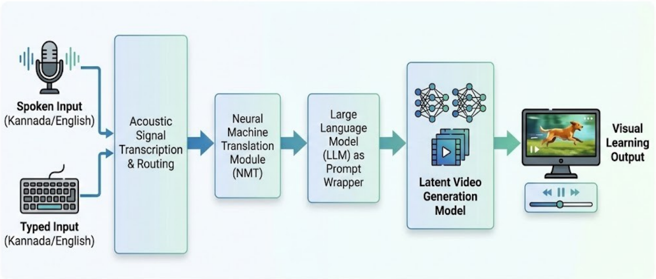
Spoken Input
(Kannada/English)
Typed Input
(Kannada/English)
Acoustic Signal Transcription & Routing
Neural Machine Translation Module (NMT)
Large Language Model (LLM) as Prompt Wrapper
Latent Video Generation Model
Visual Learning Output

Fig. 2: Detailed Proposed Educational Pipeline: The user provides a spoken or typed word (in Kannada or English) → The acoustic signal is transcribed and routed through a Neural Machine Translation Module → A Large Language Model acts as a prompt wrapper → A Latent Video Generation Model synthesizes the concept → The final Visual Learning Output is streamed to the user interface.

ceptual architecture is proposed. The proposed solution envisions a progressive web application meticulously designed for maximum simplicity, high availability, and rapid execution speed. The underlying pipeline integrates state-of-the-art natural language processing with optimized generative video diffusion models.

A. *Multimodal Input Processing and Normalization*

The user-facing interface is designed to be highly intuitive, accepting input either via a device microphone or via standardized text fields. Handling this input requires robust preprocessing:

- **Acoustic Voice Input:** The system utilizes a cloud-based automatic speech recognition application programming interface designed to capture raw audio waveforms. To accommodate the target demographic, the recognition module is configured to possess high tolerance for ambient classroom noise and speech irregularities. It features an auto-detection routing protocol to instantly determine if the incoming phonetic sequence belongs to the English or Kannada language models.
- **Textual Input:** The system allows users to manually type words, supporting distinct unicode entry modes for standard English characters and complex Kannada scripts. The raw text undergoes immediate normalization, stripping away punctuation and converting inputs to a standardized lower-case format to prevent query errors in the subsequent backend modules.

B. *Neural Translation and Automated Prompt Engineering*

Because the most advanced, open-source generative video models (which produce the highest quality outputs) are overwhelmingly trained on massive datasets of English text-video pairs, an intelligent translation and semantic expansion layer is absolutely essential for the system to function.

- **Semantic Bridging:** If the initial input is detected as Kannada (for example, “ನಾಯಿ (Nayi)”), it is immediately routed through a neural machine translation protocol optimized for Dravidian languages. This module translates the regional term into its exact English semantic equivalent (“Dog”).
- **Large Language Model Prompt Wrapping:** Raw, single-word prompts (like ”Dog”) often cause generative models to produce low-quality, abstract, or unpredictable results. To mitigate this, the translated keyword is programmatically injected into a predefined prompt template using a lightweight large language model. This prompt wrapper automatically expands the single word into a highly descriptive generation string (for example, “A cinematic, high-definition, brightly lit, educational video of a Dog playing in a grassy park, photorealistic style, clear background”). This crucial step ensures the generative model is forced to produce a clear, high-quality output rather than a blurry, confusing abstraction.

C. *Generative Video Synthesis and Delivery*

The highly engineered, descriptive text prompt is finally transmitted via application programming interface to a cloud-hosted text-to-video artificial intelligence model (utilizing architectures similar to advanced latent diffusion models).

- **Frame Synthesis:** The diffusion model synthesizes a short, continuous video clip (typically constrained to ten to fifteen seconds to minimize rendering latency) that accurately and dynamically visually represents the requested word.
- **Web Delivery:** Once rendered, the video file is compressed and streamed directly back to the user’s web application interface. This entirely automated, end-to-end process successfully creates an instant, powerful cognitive link between the specific word the child spoke or typed and the dynamic visual reality of that physical object, fundamentally transforming the vocabulary acquisition process.

VI. Conclusion

This paper provided a rigorous and comprehensive review of the rapidly evolving landscape of generative artificial intelligence, specifically examining these technologies through the critical lens of digital accessibility and special education. While the broader technology sector currently sees explosive, well-funded growth in automated entertainment, synthetic media creation, and artificial persona generation, there is a massive, largely untapped opportunity to repurpose these incredibly powerful mathematical architectures for profound social good.

The extensive literature review unequivocally highlighted that current speech-to-animation systems are fundamentally misaligned with the needs of early childhood special education. They are generally too computationally complex, overly narrative-driven, or far too specialized (such as lip-synchronizing avatars) to be utilized for basic, foundational vocabulary learning. Furthermore, the persistent lack of integrated support for regional languages continues to marginalize non-English speaking demographics.

A highly specific, urgent need was identified for a technological tool that serves as a direct visual translator for the hearing impaired—a system capable of seamlessly converting spoken or written words directly into dynamic video concepts. By intentionally shifting the technological focus away from complex story generation towards pure, unambiguous object learning, and by strictly demanding the integration of regional language support for languages like Kannada, the proposed system architecture aims to provide a highly practical, everyday learning aid.

Ultimately, the successful implementation of the proposed framework holds the potential to democratize access to essential educational information. It allows children with severe hearing impairments to independently explore, query, and profoundly understand the complex physical world around them through the universally understood language of moving video, significantly leveling the educational playing field.

References

[1] Patrick Kwon and Chen Chen, “DreamingComics: A Story Visualization Pipeline via Subject and Layout Customized Generation using Video Models,” in *arXiv preprint arXiv:2512.01686*, 2025.

[2] Zhen Bin It, Jovan Bowen Heng, and Tee Hui Teo, “Integrating Generative AI-Based Script Writing with Story Visualization: A Comprehensive Approach to Automated Narrative Creation,” in *Preprints.org*, 2025.

[3] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou, “StoryDiffusion: Consistent Self-Attention for Long-Range Image Generation,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024.

[4] Jezia Zakraoui, Moutaz Saleh, Somaya Al-Maadeed, and Jihad Mohamad Alja’am, “A Pipeline for Story Visualization from Natural Language,” in *Applied Sciences*, vol. 13, no. 8, p. 5107, 2023.

[5] Mohammad Mahdi Abootorabi, Omid Ghahroodi, Pardis Sadat Zahraei, Hossein Behzadasl, and Alire Mirrokni, “Generative AI for Character Animation: A Comprehensive Survey of Techniques, Applications, and Future Directions,” in *arXiv preprint arXiv:2504.19056*, 2025.

[6] Brayan Bernardo and Paula Costa, “A Speech-Driven Talking Head based on a Two-Stream Transformer Network,” in *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR)*, pp. 580–586, 2024.

[7] K. L. Liu and M. H. Wu, “Speech-Driven Cartoon Animation with Emotion-Controllable Face Generation,” in *IEEE Transactions on Multimedia*, vol. 26, pp. 450-462, 2024.

[8] Hyungjun Doh, Jingyu Shi, Rahul Jain, Heesoo Kim, and Karthik Ramani, “An Exploratory Study on Multimodal Generative AI in AR Storytelling,” in *arXiv preprint arXiv:2505.15973*, 2025.

[9] Sreevathsa Sree Charan, V. Srihitha, S. Palaniswamy, and S. Chethana, “Real-Time Speech-to-Text Holographic Communication for the Deaf Children and Elderly,” in *Fourth International Conference on Multimedia Processing, Communication and Information Technology (MPCIT)*, 2024.

[10] Corina-Marina Mirea, Razvan Bologa, and Andrei Toma, “Transforming Learning with Generative AI Framework: From Student Perceptions to the Design of an Educational Solution,” in *Electronics*, vol. 13, no. 12, 2024.