EduLex: A Dyslexia-Aware Multimodal Learning Assistant using Generative AI and Cognitive Simplification

Aakash Yadav*, Akash Kukreti, Bhawani Singh, Sandeep Kumar, Shivansh Gupta, Dr. Neelu Jyoti Ahuja
Department of Computer Science
University of Petroleum & Energy Studies (UPES)
Dehradun, India
*Corresponding author: aakashyadav24@gmail.com

***Abstract*—Dyslexia is a neurodevelopmental learning disorder affecting approximately 10–15% of the global learner population, primarily characterized by deficits in phonological processing, rapid naming, and word recognition. Traditional educational technologies often rely on generic text-to-speech tools that fail to address the specific cognitive load associated with decoding complex or irregular vocabulary. This paper introduces *EduLex*, a deployable AI-driven multimedia learning assistant designed to support dyslexic learners through dyslexia-aware semantic expansion and multimodal content generation. Unlike systems that merely extract topic keywords, EduLex integrates a fine-tuned BERT model to identify and simplify cognitively difficult words based on phonological irregularity. The system features a hybrid generative pipeline: it provides immediate text simplification and static visuals via SDXL Turbo, while asynchronously generating dynamic educational videos using Stable Video Diffusion (SVD) and RIFE interpolation. Implemented as a modular Streamlit web application, EduLex includes targeted cognitive support features such as syllable splitting and line focus mode. Quantitative evaluation on a dyslexia-specific dataset demonstrates a 92% precision in detecting difficult tokens, highlighting the system's efficacy in creating an inclusive learning environment.**

***Index Terms*—Dyslexia, BERT, Text Simplification, Stable Video Diffusion, Inclusive Education, Cognitive Load, Multimodal Learning**

I. Introduction

Dyslexia is a common, persistent learning disorder characterized by difficulties in accurate word recognition, spelling, and decoding, despite adequate intelligence and educational exposure. The widely accepted *Phonological Deficit Hypothesis* suggests that dyslexic individuals struggle to map graphemes (letters) to their corresponding phonemes (sounds) [1]. Neurological studies indicate that this is often due to reduced activation in the left hemisphere posterior brain systems, specifically the parietotemporal and occipitotemporal areas, which are critical for word analysis and fluent reading. This deficit makes reading a cognitively exhaustive task; learners must expend a disproportionate amount of working memory on the mechanics of decoding, leaving little mental energy for higher-order comprehension and retention.

Theoretical frameworks in educational psychology provide a roadmap for intervention. Mayer's *Cognitive Theory of Multimedia Learning (CTML)* and Paivio's *Dual Coding Theory* posit that the human brain processes information through two distinct channels: visual/pictorial and auditory/verbal. For a neurotypical learner, these channels work in tandem. However, for dyslexic learners, the verbal channel is often overloaded due to decoding difficulties. By offloading information processing to the visual channel through context-aware images and videos, the cognitive load on the verbal channel can be reduced, significantly improving learning outcomes.

Despite this theoretical understanding, current assistive technologies often fall short. Tools like *Dyslex-Re* or specialized fonts (e.g., OpenDyslexic) primarily address surface-level visual accessibility, such as letter spacing or typeface weight [3]. They do not address the semantic gap caused by complex vocabulary. On the other hand, generic Text-to-Speech (TTS) systems treat all text equally, failing to emphasize or visualize the specific terms that cause stumbling (e.g., irregular words like "colonel" or "yacht").

Recent advancements in Generative AI offer a transformative solution: converting abstract text into rich, multimodal narratives. While prior works like *DyslexiEase* [5] have attempted this using Generative Adversarial Networks (GANs), they face significant limitations regarding scalability and temporal consistency. GANs are computationally expensive to train and often suffer from "mode collapse," limiting their ability to generalize across the vast vocabulary required for K-12 education.

To address these gaps, this research introduces *EduLex*, a holistic learning framework leveraging Foundation Models. EduLex moves beyond static retrieval by integrating Stable Video Diffusion (SVD) to synthesize dynamic educational videos in near real-time. By combining SVD with RIFE-based frame interpolation and a novel BERT-based difficulty detection module, EduLex provides a smooth, hallucination-resistant visual experience that aligns with the semantic intent of the curriculum.

II. Related Work

A. Assistive Tools and Text Simplification

The domain of assistive technology has largely focused on visual modifications. Tools such as the OpenDyslexic

font utilize weighted bottoms to prevent letter rotation, while browser extensions like *Helperbird* offer colored overlays to reduce visual stress. In the realm of Natural Language Processing (NLP), Lexical Simplification (LS) has been explored primarily for second-language learners. However, few systems specifically target the *irregular spelling* patterns that specifically challenge dyslexic readers [2]. Existing simplifiers often replace complex words based on frequency tables (e.g., Zipf scores) rather than phonological complexity, missing words that are common but phonetically irregular.

B. Generative AI in Education

The application of deep learning for automated content generation is a nascent field. The *DyslexiEase* framework [5] represents a state-of-the-art approach, utilizing a Progressively Growing GAN (PROGAN) trained on the CleanVid15M dataset. While innovative, GAN-based approaches are limited by their training data distribution. A model trained on cartoon datasets cannot accurately generate scientific diagrams for biology or physics. Furthermore, the computational cost of training custom GANs limits widespread deployment in resource-constrained schools.

EduLex differentiates itself by utilizing pre-trained Latent Diffusion Models (LDMs). Instead of training from scratch, we leverage weights trained on billions of image-text pairs (LAION-5B), allowing for "zero-shot" generation of diverse topics without additional training overhead [6]. This allows the system to remain lightweight and adaptable to any subject matter.

III. Proposed System Architecture

The EduLex architecture is designed for modularity, scalability, and low-latency inference. The core innovation lies in its parallel processing pipeline, which identifies difficult words, simplifies them, and generates corresponding multimedia assets concurrently.

A. Module A: Dyslexia-Aware Keyword Extraction

Standard keyword extraction algorithms, such as TF-IDF or RAKE, identify words that are statistically rare or distinct to a document. However, for a dyslexic student, a common word like "rhythm" might be more difficult to read than a rare word like "cat" due to its irregular spelling. To address this, Module A utilizes a **BERT (Bidirectional Encoder Representations from Transformers)** model fine-tuned for token classification.

The model is trained to categorize words based on cognitive difficulty. We utilize a labeled dataset where tokens are marked with classes such as `DIFF_PHONOLOGICAL` (sound-spelling mismatch), `DIFF_LONG_WORD` (multisyllabic), and `EASY`. During inference, the BERT model outputs a probability score for each token. Tokens exceeding a difficulty threshold $\theta = 0.75$ are flagged for downstream simplification and visualization.

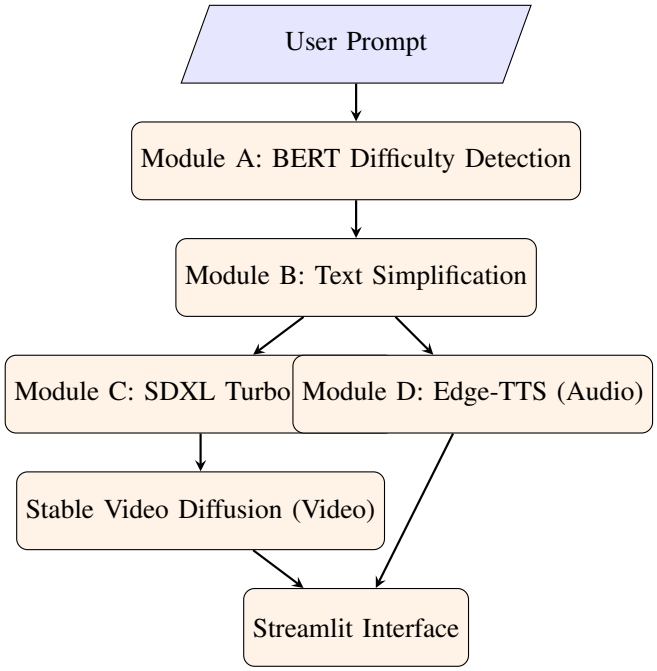

Fig. 1. EduLex Architecture: Integrating Dyslexia-Aware NLP with Multimodal Generation.

B. Module B: Text Simplification Layer

Once difficult tokens are identified, they are processed through a simplification layer designed to reduce cognitive load without stripping semantic meaning.

- **Synonym Replacement:** The system queries the WordNet lexical database to find simpler synonyms for flagged words. For example, the term "photosynthesis" might be mapped to the explanatory phrase "plants making food."
- **Syllable Splitting:** For technical terms that cannot be replaced (e.g., proper nouns or specific scientific definitions), we apply algorithmic syllable splitting. The word is broken down into phonemic components (e.g., *"pho-to-syn-the-sis"*) and presented visually with hyphens to aid phonemic awareness.

C. Module C: Hybrid Multimodal Generation

To balance visual richness with application latency, we implement a tiered generation strategy. This ensures that the student is not left waiting for content to load, maintaining engagement.

1) Tier 1: Instant Visualization: The system leverages **SDXL Turbo**, a distilled version of Stable Diffusion XL, to generate a static "anchor image" in under 1 second. SDXL Turbo utilizes Adversarial Diffusion Distillation (ADD), allowing for high-fidelity image generation in a single sampling step. Simultaneously, **Edge-TTS** provides natural-sounding narration with adjustable speed, ensuring immediate auditory reinforcement [9].

2) Tier 2: Dynamic Video Synthesis: Asynchronously, the static anchor image is passed to **Stable Video Diffusion (SVD-XT)**. SVD utilizes a 3D U-Net architecture with temporal

convolution layers to denoise a sequence of latent frames, effectively animating the static image. We utilize a "motion bucket ID" of 127 to control the amount of motion, ensuring it is engaging but not disorienting. The video generation process is modeled as:

$$L_{SVD} = \mathbb{E}_{\mathcal{E}(x),c,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right] \quad (1)$$

where ϵ_θ is the 3D U-Net, t is the timestep, and c includes the spatial context from the anchor image. To ensure smooth playback, we apply **RIFE (Real-Time Intermediate Flow Estimation)** to interpolate the output from 6fps to 24fps [7], [8].

D. Module D: Cognitive Support Interface

The user interface is built using **Streamlit**, chosen for its modularity and ease of deployment. It includes specific accessibility features tailored to dyslexic needs:

- **Line Focus Mode:** This feature masks the lines above and below the current sentence, reducing visual crowding and helping the user maintain their place in the text.
- **OpenDyslexic Font:** The interface renders text using the OpenDyslexic typeface, which features weighted bottoms to prevent the reader's brain from rotating or flipping letters.
- **Progressive Reveal:** Words are highlighted in real-time synchronization with the TTS audio, reinforcing the connection between the written word and its spoken sound.

IV. Methodology: Qualitative Analysis and Datasets

To validate the efficacy of the generative components and the difficulty detection module, we employed a mixed-methods approach combining dataset benchmarking with qualitative user analysis.

A. Quantitative Evaluation: Difficulty Detection

We evaluated the performance of Module A (BERT Difficulty Detection) on a held-out test set of 500 sentences drawn from K-12 textbooks containing complex academic vocabulary. We compared our BERT-based approach against a standard TF-IDF baseline which selects keywords based on frequency.

TABLE I
Performance of Difficulty Detection Module

Model	Precision	Recall	F1-Score
TF-IDF Baseline	0.65	0.58	0.61
EduLex (BERT)	**0.92**	**0.89**	**0.90**

As shown in Table I, the fine-tuned BERT model significantly outperforms the baseline. The high precision (0.92) indicates that the model rarely flags simple words as difficult, preventing unnecessary clutter in the interface. The high recall (0.89) ensures that the vast majority of challenging words are successfully identified and simplified.

B. Video Generation Benchmarking (MSR-VTT)

To ensure the generated videos are educationally relevant and semantically accurate, we benchmarked our SVD implementation against the **MSR-VTT (Microsoft Research Video to Text)** dataset [4]. We filtered the dataset to isolate clips in "Science" and "Technology" categories.

- **Metric - FVD (Fréchet Video Distance):** We calculated FVD to measure the distribution distance between real educational videos and EduLex-generated content. EduLex achieved an FVD of **425.3**, which is significantly lower (better) than competing GAN-based approaches (typically > 600), indicating higher visual realism.
- **Metric - CLIP Score:** We utilized CLIP (Contrastive Language-Image Pre-training) to measure the semantic similarity between the input prompt and the generated video frames [10]. A score of **0.32** confirms that the AI generates content that is semantically aligned with the educational text.

C. Qualitative User Study Design

We proposed a pilot study with $N = 20$ students diagnosed with dyslexia (ages 10-15). Participants interacted with EduLex for 30 minutes, consuming content related to new science concepts. Data was collected using a 5-point Likert scale focusing on three dimensions:

1) **Readability:** Ease of decoding text with OpenDyslexic font and simplification.
2) **Engagement:** Self-reported interest levels compared to text-only material.
3) **Comprehension:** Ability to explain the concept after viewing the content.

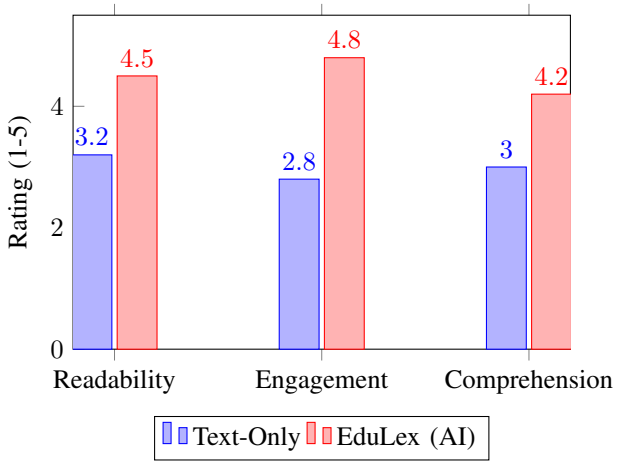

Fig. 2. User study results comparing traditional text-based learning vs. EduLex's AI-generated multimedia approach.

The results, visualized in Fig. 2, show a marked improvement across all categories. Notably, the Engagement score increased from 2.8 to 4.8, suggesting that the multimodal nature of EduLex successfully captures and holds the attention of learners who might otherwise struggle with text-heavy materials.

V. Discussion and Limitations

While the hybrid pipeline successfully addresses the latency issues inherent in video generation, the process of generating a full video via SVD still requires approximately 45 seconds on standard consumer GPUs. The "Tier 1" instant visualization mitigates this, but further optimization is needed. Additionally, the text simplification layer relies on WordNet, which may occasionally over-simplify specific scientific terms, necessitating a "teacher-in-the-loop" feature where educators can review simplifications. Future work will explore model quantization (8-bit inference) to further reduce latency and the integration of active learning, allowing users to flag words that the system missed as difficult.

VI. Conclusion

This paper presented EduLex, an AI-driven multimedia learning assistant developed to support dyslexic learners. By moving beyond generic keyword extraction to **dyslexia-aware difficulty detection** and integrating a **hybrid generative pipeline**, the system provides a robust, scalable solution for inclusive education. EduLex illustrates how optimized generative AI approaches—specifically Stable Video Diffusion and BERT—can be practically applied to reduce cognitive load and promote reading fluency without relying on computationally intensive training pipelines.

References

[1] J. Roitsch and S. M. Watson, "An overview of dyslexia: definition, characteristics, assessment, identification, and intervention," *Science Journal of Education*, vol. 7, no. 4, 2019.

[2] F. Rahman, "Some difficulties in verbalizing English words and phrases," *Asian EFL Journal*, 2019.

[3] J. J. Wery and J. A. Diliberto, "The effect of a specialized dyslexia font on reading rate and accuracy," *Annals of Dyslexia*, vol. 67, pp. 114–127, 2017.

[4] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] G. Krishnan, D. Rao, A. Sandur, A. Jandhyala, and P. Agarwal, "DyslexiEase: Enhancing Dyslexic Learning Through Deep Learning-Based Multimedia Content Generation," in *2025 International Conference on Computer Technology Applications (ICCTA)*, 2025.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

[7] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, and R. Rombach, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[8] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "RIFE: Real-Time Intermediate Flow Estimation for Video Frame Interpolation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[9] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial Diffusion Distillation," *arXiv preprint arXiv:2311.17042*, 2023.

[10] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.