# Multilingual Speech-to-Video Generation for Hearing-Impaired Object Learning

## Simplifying Visual Learning with Generative AI

Sandesh PY

Dept. of Computer Science & Engineering
Vidyavardhaka College of Engineering, Mysuru

# Introduction

- **The Core Issue:** Hearing-impaired children struggle to learn vocabulary because they lack auditory reinforcement.
- **Current Limitation:** Traditional tools (flashcards) are static and boring. Existing AI tools (talking heads) only show faces, not objects.
- **Our Solution:** An AI-powered "Visual Dictionary" that instantly converts spoken or typed words into dynamic videos.
- **Key Feature:** Supports both **English** and **Kannada** to help local students.

# Motivation

## Why is this necessary?

1. **Cognitive Gap:** A child needs to "see" the verb "run" to understand it; a static picture isn't enough.
2. **Language Barrier:** Most AI tools are English-only. Rural students in Karnataka need tools in their mother tongue.
3. **Accessibility:** Complex apps require expensive AR/VR gear. We need a simple web tool that works on any phone or laptop.

# Project Objectives

- **Objective 1:** Develop a web-based "Visual Learning Assistant" for hearing-impaired children.
- **Objective 2:** Implement **Multimodal Input** (Voice  Text) to support different learning stages.
- **Objective 3:** Integrate **Kannada Language Support** using Neural Machine Translation.
- **Objective 4:** Utilize Generative AI (Latent Diffusion) to create clear, unambiguous videos of objects (e.g., "Apple", "Car").

# Literature Survey  Gap Analysis

| Approach | Focus | Limitation |
|---|---|---|
| StoryDiffusion | Long narrative consistency | Too slow; complex scenes confuse learners. |
| Talking Heads | Lip-syncing avatars | Only shows faces; cannot teach "Elephant". |
| Static Flashcards | Basic vocabulary | No motion; less engaging. |

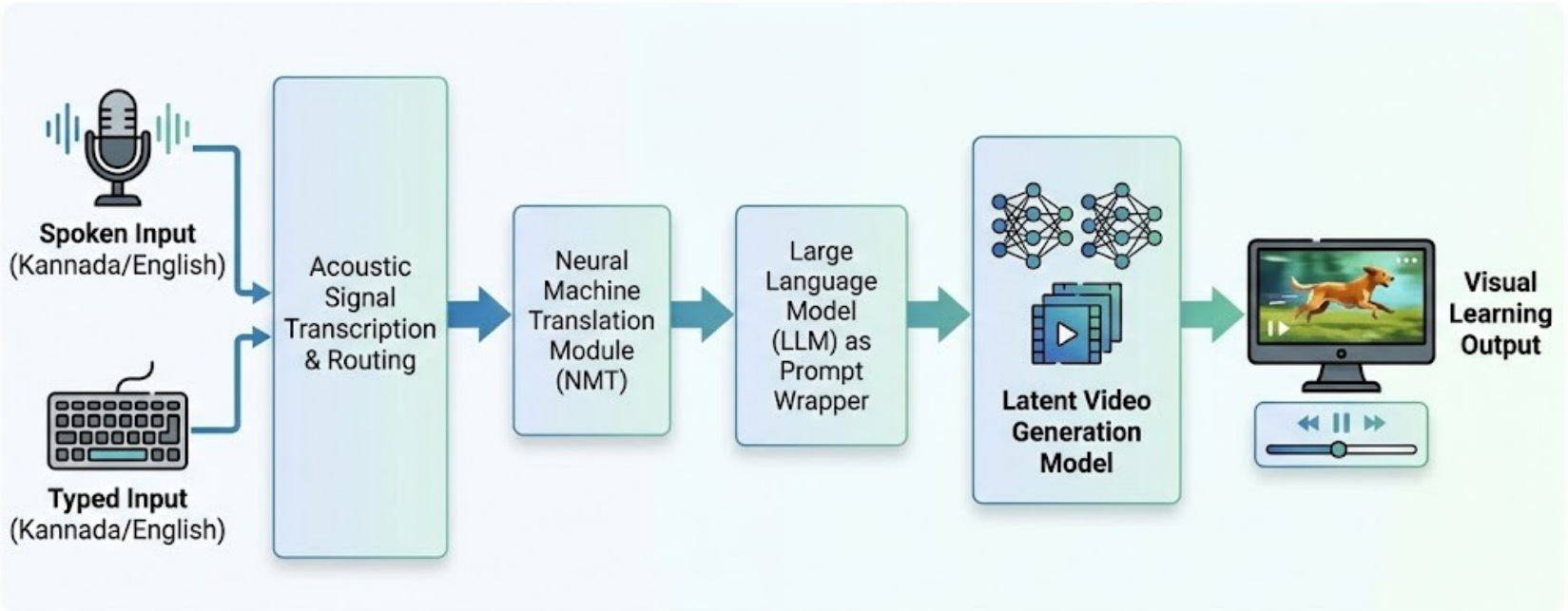
**Identified Gap:** There is no simple, fast, and multilingual tool designed specifically for *object visualization*.

# Proposed Methodology

The system follows a 4-step linear pipeline:

1. **Input:** User speaks "Nayi" (Kannada word for Dog) or types "Dog".
2. **Processing:**
   - Speech-to-Text converts audio.
   - Translation Module converts Kannada $\rightarrow$ English ("Dog").
3. **Enhancement:** LLM expands "Dog" $\rightarrow$ "A cinematic video of a cute dog running in a garden."
4. **Generation:** Text-to-Video Model synthesizes the final video clip.

# System Architecture




User Interface → Flask Backend → Google Colab (AI Engine) → Video Output

# Implementation Tech Stack

**Frontend:**

- HTML5, CSS3, JavaScript
- Simple, child-friendly UI

**Backend:**

- Python (Flask)
- SQLite Database

**AI Core (Colab):**

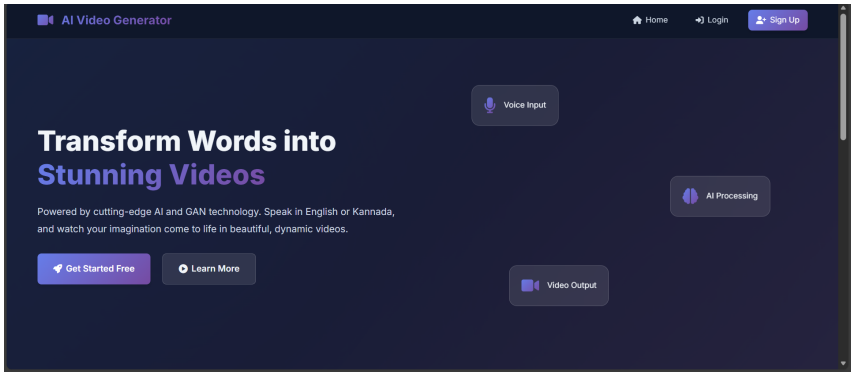
- **Model:** Zeroscope (Latent Video Diffusion)
- **Translation:** Google Translate API
- **Voice:** SpeechRecognition Lib

# Results Discussion

- Successfully generates 3-second videos for standard objects.
- **Latency:** Average generation time is 60 seconds (using Cloud GPU).
- **Accuracy:** High accuracy for common nouns (Animals, Vehicles, Fruits).
- **Language:** Successfully handles Kannada voice input and maps it to the correct visual.

# User Interface Results

## Web Interface




## Dashboard Output

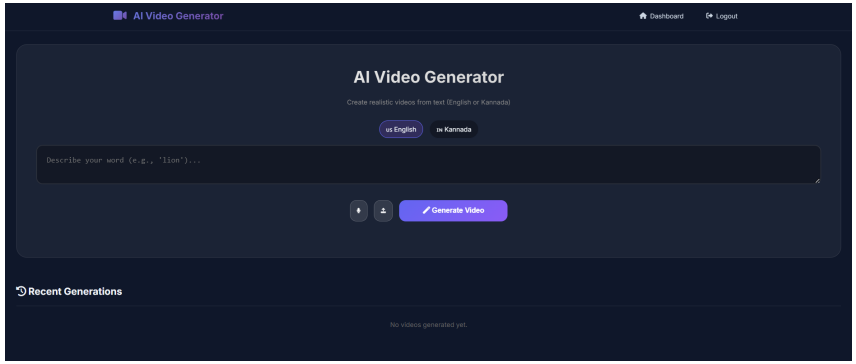



Figure: The Left image shows the web interface ; The Right image shows the multilingual input screen.

# Generated Video Output

## Final Generated Video Visualization

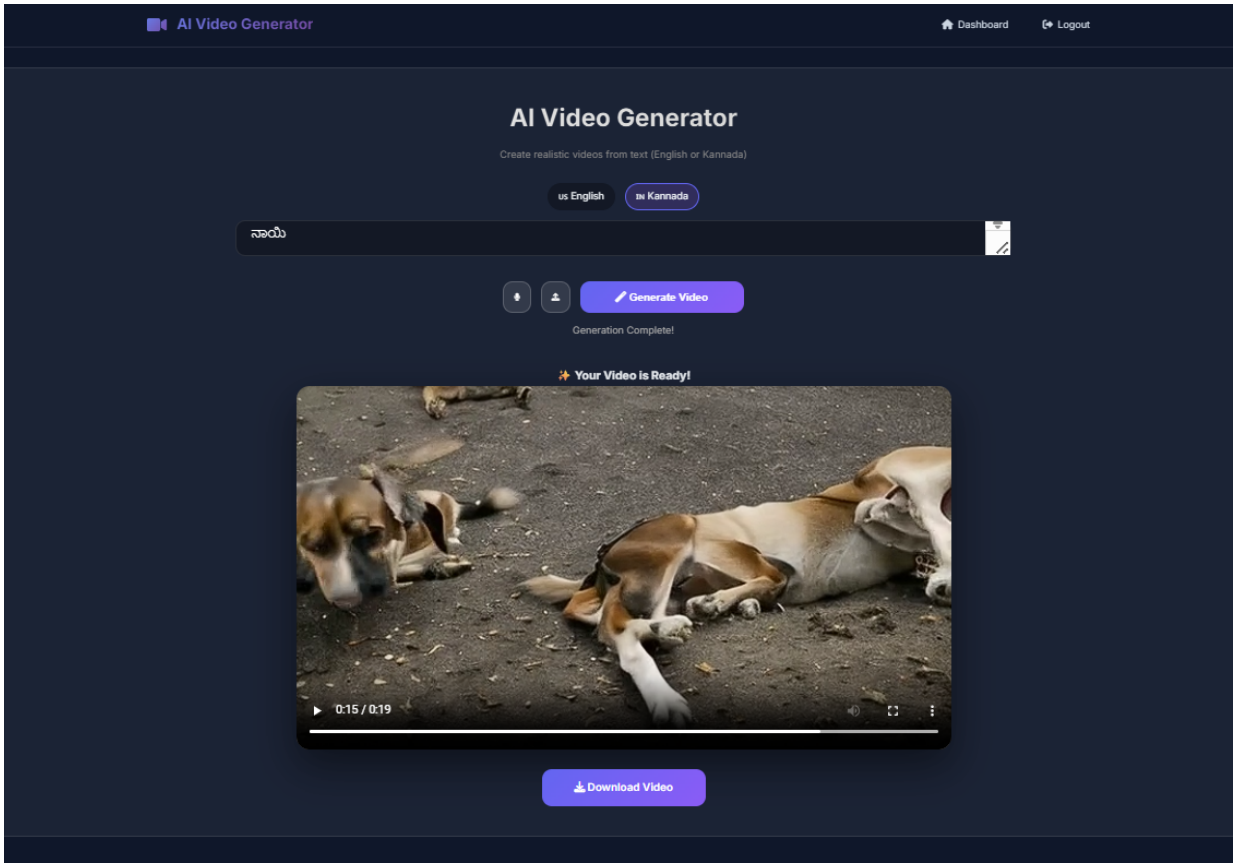



Figure: Screenshot of the generated video playback and download options.

# Conclusion Future Scope

**Conclusion:**

- We successfully built a prototype that bridges the gap between text and visual understanding for special education.
- The inclusion of regional language support makes it accessible to a wider demographic.

**Future Scope:**

- Reduce generation time to real-time ($<$5 seconds).
- Add "Gamification" (Quizzes) to test the child's learning.
- Support specialized Sign Language video generation.

# Thank You!