NeuralPlay

Technical Specification for an AI-Powered Video Player

Prepared by Gbohunmi Oredipe for NeuralPlay

Date: October 2025

1. Executive Summary

NeuralPlay is an advanced, AI-powered video player application designed for Windows operating systems. The system integrates machine learning and computer vision models to enhance user experience while watching videos. NeuralPlay offers automatic speech-to-text caption generation, intelligent transcript search, scene segmentation with AI-based summaries, object and person detection, facial emotion recognition, interactive Q&A; about video content, and voice command control. The application operates both online and offline, providing flexibility through a hybrid model pipeline that uses local lightweight models for minimal operations and cloud-based APIs for enhanced accuracy and performance.

2. System Overview

NeuralPlay is structured as a hybrid desktop application that combines multimedia playback with artificial intelligence modules. The software utilises Electron for its desktop interface, Python for backend processing, and a collection of pre-trained AI models for feature implementation. The system is modular, allowing future updates and expansion without restructuring the entire architecture.

3. Functional Requirements

1. Play and control local video files (MP4, MKV, AVI).
2. Generate captions from video audio using AI speech recognition.
3. Display synchronised transcripts alongside playback.
4. Detect and segment scenes, generating automatic summaries.
5. Detect visible objects and persons using computer vision.
6. Identify and classify emotions on detected faces.
7. Enable users to query video content through text-based Q&A.;
8. Accept basic voice commands for playback and navigation.
9. Allow both offline processing and cloud-assisted enhancement.
10. Store metadata, captions, and analysis locally for reuse.

4. Non-functional Requirements

- The application must provide real-time or near real-time response for AI-assisted functions.
- The system must operate efficiently on Windows machines with minimal resource consumption.
- Data must be securely stored locally, with user consent required for cloud operations.

- The interface must remain responsive even during AI processing.
- The architecture must be modular and maintainable, supporting model and API upgrades.

5. Detailed AI Feature Descriptions

• *Speech-to-Text Captions*

Uses OpenAI Whisper or Whisper.cpp for automatic transcription of spoken content. Audio is extracted using FFmpeg and processed into timestamped captions.

• *Smart Transcript and Search*

Stores complete video transcripts in a searchable database. Users can locate specific dialogue and jump directly to corresponding timestamps.

• *Scene Detection and AI Summaries*

Employs OpenCV and FFmpeg to detect scene boundaries based on colour histograms or frame differences. Summaries are generated via GPT-based natural language models.

• *Ask-the-Video (Q&A;)*

Allows users to ask natural language questions about video content. Relevant transcript sections are retrieved using semantic search, and responses are generated through LLMs such as GPT or Mistral.

• *Object and Person Detection*

Implements YOLOv8 or YOLOv10 pretrained models to detect and label visible entities in frames. Results can be overlaid visually or used for search indexing.

• *Emotion and Face Recognition*

Integrates facial analysis models (e.g., DeepFace) to recognise emotional states and expressions.

• *Voice Command Control*

Supports basic control commands like play, pause, skip, and volume adjustment through speech recognition.

6. System Architecture and Workflow

The architecture of NeuralPlay is divided into four major layers: frontend, backend, AI engine, and storage. The frontend, built using Electron and React, provides the user interface. The backend, implemented in Python using FastAPI, acts as the bridge between the interface and AI modules. The AI engine houses integrated models for speech recognition, computer vision, and natural language processing. Finally, the storage layer, powered by SQLite, maintains metadata, captions, transcripts, and user preferences.

7. Data Management and Storage

NeuralPlay stores video metadata, transcripts, scene data, and AI results locally using SQLite. Each video is assigned a unique ID linking captions, transcripts, and AI analyses. The system also maintains a local cache for faster retrieval of previously processed videos. User data is kept private, and any cloud interaction requires explicit permission.

8. User Interface Specification

The user interface includes a main video panel, a side panel for AI-generated data (captions, transcript, summaries, detections), and an optional voice control toolbar. Users can toggle features, search within transcripts, and interact with AI outputs in real time. The design prioritises clarity, accessibility, and performance.

9. AI Model Integration (Local and Online)

NeuralPlay operates in a hybrid mode. For offline use, lightweight models such as Whisper.cpp and YOLOv8-tiny handle basic operations. When connected to the internet, the system may leverage cloud-based APIs like OpenAI Whisper or GPT for improved accuracy and faster processing. The transition between modes is automatic, based on network availability and user settings.

10. Development Phases and Implementation Plan

1. Phase 1: Core video player, transcription, and local storage.
2. Phase 2: Transcript search, scene detection, and AI summarisation.
3. Phase 3: Object detection, emotion recognition, and visual overlays.
4. Phase 4: Voice command control, hybrid mode refinement, and UI enhancement.

11. Future Scalability

Future development of NeuralPlay will focus on cross-platform deployment (macOS, Android), integration with streaming services, support for real-time subtitle translation, and plugin-based AI feature expansion. The modular nature of the architecture ensures that new models or APIs can be added without disrupting existing functionality.

12. Conclusion

NeuralPlay represents a significant evolution in media playback technology by merging artificial intelligence with traditional video player design. Through its comprehensive feature set, hybrid processing capability, and user-centric design, it aims to redefine how users interact with video content.