Product Requirements Document (PRD)

1. Overview

A cross-functional initiative to build a **local, video-only Retrieval-Augmented Generation (RAG) assistant** that answers natural-language questions about a curated video library and returns (1) an answer that cites precise timestamps and (2) auto-generated clips of the referenced moments. The system targets small research teams, educators, and content creators who need rapid, citation-accurate insights from long video assets without uploading data to external clouds.

2. Vision & Value Proposition

> **"Surface the exact 30 seconds you need from hours of footage—in one question."**

- **Time-savings**: Cut search time from hours of manual scrubbing to seconds.
- **Privacy-first**: Entirely local processing, no data leaves the user's machine.
- **Seamless evidence**: Answers always backed by verifiable timestamps and playable clips.

3. Goals & Success Metrics

Goal	Metric	Target (MVP)
Accurate retrieval	Answer cites a correct timestamp (±5 s)	≥ 85 % of test queries
User satisfaction	SUS score	≥ 80
Clip generation speed	From query to playable clip	≤ 30 s on dev laptop
Local footprint	Disk usage (models + DB)	≤ 8 GB

4. Personas

Persona	Need
Research Analyst	Quickly locate spoken claims in interview recordings for fact-checking.
Educator	Pull short illustrative segments from lectures for slides.
Content Creator	Identify and clip highlight moments for social media.

5. User Stories (Top Priority)

1. **Ask & Clip**
   As a user, I can ask a question in plain English and receive an answer that cites where in the video the answer comes from, so that I can immediately watch that moment.

2. **Evidence Link**
   As a user, I can click a cited timestamp in the answer and the video begins playing from that exact point.
3. **Bulk Add Videos**
   As a user, I can drop multiple local videos into the app and start querying them after processing completes.

6. Functional Requirements

#	Requirement
F-1	System shall extract audio and 10-second-granular transcripts using Whisper.
F-2	System shall sample one frame every 10 s and store its time range.
F-3	All segments (audio & frame) shall be embedded with CLIP and stored in a local vector DB (Chroma).
F-4	Upon query, system shall retrieve the top-10 semantically nearest segments via cosine similarity.
F-5	System shall prompt an API-based LLM to generate an answer that references video IDs and timestamps.
F-6	System shall parse cited timestamps and auto-export the corresponding clips (H.264).
F-7	UI shall list the answer and render the video clips with play controls.

7. Non-Functional Requirements

- **Privacy**: No internet upload of video or transcript data.
- **Performance**: Entire query→clip loop ≤ 30 s on a modern CPU laptop.
- **Extensibility**: Pipeline phases callable independently via CLI or API for automation.
- **Observability**: Logs and metrics for each phase (processing time, token usage, clip count).

8. Out of Scope (MVP)

- Scene-change detection
- Multi-modal (images, PDFs) ingestion
- Mobile UI

9. Technical Approach (Summary for Stakeholders)

1. **Six-Phase Modular Pipeline** (see accompanying Development Plan document). Each phase exposes a clear interface so teams can work in parallel and swap implementations later.
2. **Local-only stack**: Python, FFmpeg, Whisper (CPU), OpenClip, ChromaDB, FastAPI, LangChain.
3. **Pure CLIP embeddings** for both text and frames to keep a single search space and simplify retrieval.
4. **Timestamp-first data model**: Every stored vector carries start/end metadata, ensuring traceability from query → answer → clip.

10. Risks & Mitigations

Risk	Likelihood	Impact	Mitigation
Long transcription time on CPU	Med	Med	Pre-process overnight; option to drop to Whisper `small` model.
CLIP text encoder retrieval quality	Med	Med	Future upgrade path: dual-embedding with text-native model + reranker.
LLM cost per query	Low	Med	Cache embeddings & sources; limit context to 6 chunks.

11. Milestones (Executive View)

Date (Week)	Milestone
W1	Audio & frame extraction prototypes ready
W2	Local Chroma vector DB populated with test video
W3	Retrieval API returns top-10 segments
W4	First end-to-end answer incl. timestamps (no clip)
W5	Auto-clipping integrated, UX click-to-play demo
W6	Stakeholder review; decide on Post-MVP features

12. Stakeholders & Roles

- **Product Owner** – KR x OP (defines requirements, accepts features)
- **Tech Lead** – Backend 1 (architectural decisions, code review)
- **ML Lead** – ML 1 (model selection, embedding quality)
- **DevOps** – DevOps 1 (CI/CD, containerisation)
- **QA Lead** – Backend 2 (test frameworks, performance benchmarks)

13. Open Questions

1. Do we need a lightweight desktop UI or is a browser-based front-end sufficient?
2. Shall we bundle Whisper weights in installer or require separate download?
3. What legal/licensing constraints exist for distribution of CLIP weights?

This PRD is the stakeholder-friendly companion to the engineering Development Plan. Updates tracked via version-controlled docs.