

Video RAG Architecture Checklist

Fora Soft Learn · AI for Video Engineering

1 · Do you even need RAG? (vs a long-context model)

- ☐ One moderate video, a few questions: skip RAG. Send the whole thing to a long-context model. Simpler.
- ☐ Archive too BIG to fit: 1 hour of video $\approx$ 947,000 tokens ($\sim$ 263/sec). A 100-hour archive does not fit. Use RAG.
- ☐ Too COSTLY to re-read: long context re-reads the video on every query ($\sim$ \$2.37/hour at Gemini 2.5 Pro). RAG indexes once, then reads chunks.
- ☐ Accuracy: models attend unevenly past $\sim$ 400k tokens. A short retrieved context is more accurate.

2 · Pick the retrieval design (cheat-sheet)

Text grounding	ASR + OCR + scene captions $\rightarrow$ text RAG	cheapest, mature; lossy on visuals
Visual embeddings	embed frames/clips, multimodal model	heavier; pure-visual meaning survives
Hybrid	text for the bulk + visual for visual-only queries	most production systems
Rule of thumb	mostly SAID/on-screen $\rightarrow$ text mostly SEEN $\rightarrow$ visual both $\rightarrow$ hybrid	

3 · Ingestion — runs ONCE per video (carries the cost)

- ☐ Chunk by SCENE/speaker/topic, never by a fixed clock interval. Let chunks overlap slightly so boundary moments survive.
- ☐ Frame-select before processing: downsample 60 $\rightarrow$ 4 fps, then keep keyframes only ($\sim$ 90 $\times$ fewer frames).
- ☐ Extract: ASR (with timestamps) + OCR + VLM scene captions, or visual embeddings — whichever design you picked.
- ☐ Index every chunk's embedding in a vector DB (Pinecone/Weaviate/Milvus/Qdrant) WITH timestamps + source metadata.

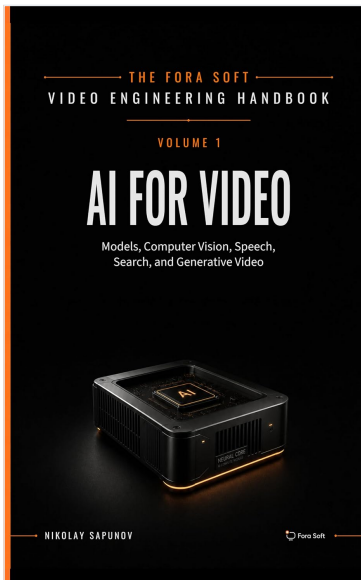
4 · Query — runs every question (keep it cheap + trustworthy)

- ☐ Embed the question to the same meaning map; fetch top-k (3-10) nearest chunks from the vector DB.
- ☐ Rerank the short candidate list to the best 2-3 before generation — cheap, because the list is already short.
- ☐ Generate from retrieved chunks ONLY, and surface cited timestamps + source clips so a human can verify in one click.
- ☐ Cost shape: pay once to index; each question reads a few thousand tokens ($<$ 1 cent). The gap widens with scale.

Companion to: 'Video RAG: Multimodal RAG Over A Video Archive'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/video-rag-multimodal-rag-over-video-archive

Fora Soft · www.forasoft.com



GO DEEPER · VOLUME 6 OF 7

AI for Video: A Complete Guide to Intelligence in Video Products

From Models and Inference to Computer Vision, Speech, Search, and Generative Video

By Nikolay Sapunov, CEO at Fora Soft · 1,444 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6TKPZ1

This sheet decides between long context and RAG, then sets chunking, frame selection and indexing for a video archive. Volume 6 covers search and RAG over video archives in depth, together with speech recognition, OCR and the vision models that produce the text you index.

WHAT YOU WILL FIND INSIDE

- Search and RAG over a video archive
- ASR, diarization and timestamps
- OCR and scene understanding
- Model artifacts and deployment topologies

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Want people to ask questions of your video archive and get cited clips back?

We will pick text grounding or visual embeddings for your content, build the index, and return answers with timestamps a reviewer can check. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)