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 ~= 947,000 tokens (~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 (~$2.37/hour at Gemini 2.5 Pro). RAG indexes once, then reads chunks.
- [ ] Accuracy: models attend unevenly past ~400k tokens. A short retrieved context is more accurate.

2 · Pick the retrieval design (cheat-sheet)

```
Text grounding    : ASR + OCR + scene captions -> 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 -> text | mostly SEEN -> visual | both -> 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->4 fps, then keep keyframes only (~90x 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.