AI Video Summarizer Build Sheet

Fora Soft Learn · AI for Video Engineering

1 - Three engines (pick by what your video means)

Engine	What it sees	Cost	Best fit
Transcript-first	reads captions only	<$0.01 / 30-min	talking-head, lectures, podcasts.
Hybrid	transcript + sampled keyframes	~$0.05-0.25	webinars, tutorials, screen-share.
Native multimodal	watches frames + hears audio	~$0.25-1.35	demos, sports, surveillance.

Deciding question: does the meaning live in the WORDS, or in the PICTURE?

2 - The five-stage pipeline (every engine shares it)

1 Acquire words	existing captions (free) OR run speech-to-text (Whisper ~10% WER).
2 Clean & segment	stitch caption cues into sentences; chunk (~10k chars / 1k overlap).
3 Summarize	stuff / map-reduce / refine (see box 3).
4 Shape output	bullets, chapter markers, JSON; attach timestamps (MM:SS).
5 Check	faithfulness gate - verify every claim against the source.

3 - Summarize a long transcript + the cost arithmetic

Map-reduce	parallel chunks then combine - fast; can't link distant sections.
Refine	running summary carried forward - reasons across; sequential, slower.
Stuffing	whole transcript in one long-context prompt - default for one-shot summaries.

30-min video: transcript ~6,000 tokens = <$0.01 | multimodal 1 FPS ~300 tok/sec

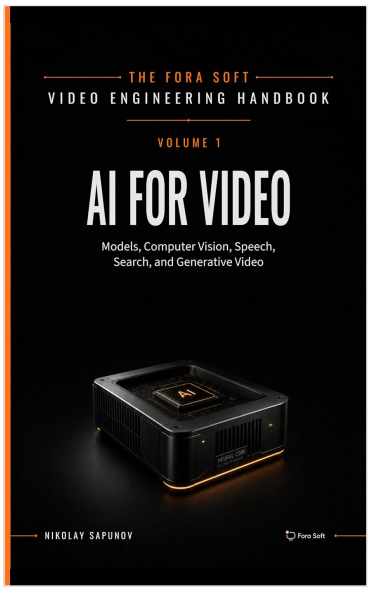
-> 1,800s x 300 = 540,000 tokens ~ $1.35 (full-res) | low-res ~100 tok/sec ~ $0.23.

4 - Build vs buy (rent the model, own the pipeline)

Rent APIs	frontier model + hosted STT - fastest, pay per use; wins at low/spiky volume.
Self-host	Whisper + open LLM - fixed cost; wins at high steady volume or strict privacy.

5 - Ship gate - quality + legal (context, not legal advice)

- [ ] Pick the engine by content, not habit: use frames only when the meaning is visual.
- [ ] Ground every claim and every timestamp to the transcript line that supports it - let users click to verify.
- [ ] Score faithfulness with an LLM-as-judge, not word-overlap (ROUGE) - hallucination is the dangerous error.
- [ ] Cache long videos you re-query; compute the rent-vs-self-host crossover before committing.
- [ ] Access captions through sanctioned API paths only - most free transcript scrapers forbid commercial use.



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 build sheet picks a summarizer engine by whether meaning lives in words or pictures, then walks the pipeline, token arithmetic and ship gate. Volume 6 covers the parts in depth: speech recognition, multimodal models over frames, search and RAG over a video archive, and the ML basics for choosing models.

WHAT YOU WILL FIND INSIDE

- ASR quality and transcript cleanup
- RAG and search over long video archives
- Multimodal models that read frames
- Choosing and evaluating models for a feature

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

Planning a video summarizer for lectures, meetings or a media archive?

We will choose the engine for your content, run the token math on your hours of video and plan a faithfulness check before launch. 250+ video products shipped since 2005.

Book a 30-min call → **WhatsApp →** **Email us →**