

Async Archive-Review Agent - Scoping & Durability

Fora Soft Learn · AI for Video Engineering

1 - The review loop, run as a fleet (the agent loop, pointed at a whole archive)

- Archive -> work queue -> N parallel workers, each: perceive -> reason -> act -> observe -> write
- Fan-out: 1,000 workers drain the line 1,000x faster. No clock, no single prompt.
- Confident verdicts -> catalog; low-confidence / severe -> human review queue (escalate by exception).

2 - The cost funnel (cheapest tools first; the expensive reader sees only what survives)

Cheap filter	shot-change + transcript; drop dead air	(runs on everything)
Keyframe sampler	pick the few frames per shot that matter	(low cost)
Reader (VLM)	describe the sampled frames	(HIGH cost - keep it last)
Policy / judge	description -> structured verdict + score	(medium)
Catalog / escalate	store the confident; send the rest to people	(low)

3 - The no-clock economics (nothing is waiting, so optimize cost, not speed)

- Batch APIs: OpenAI / Anthropic / Gemini all 50% off for a 24-hour turnaround. Real-time can't use them.
- Funnel math: 18,000 frames/video -> ~40 sampled = ~450x less reader work; then batch halves what's left.
- Spot / preemptible compute: cheap and interruptible - safe ONLY if the job survives interruption (see 4).

4 - The durability spine (a crash at item 80,000 resumes at 80,001, never restarts)

CHECKPOINT	record each done item before taking the next.
IDEMPOTENCY	key = video-N + model + policy version; skip what's already done (or you double-bill).
RETRY + DLQ	retry transient fails; a poison item -> dead-letter queue -> one human, not the whole job.
ORCHESTRATE	durable engine (e.g. Temporal) replays its event log and resumes mid-job.

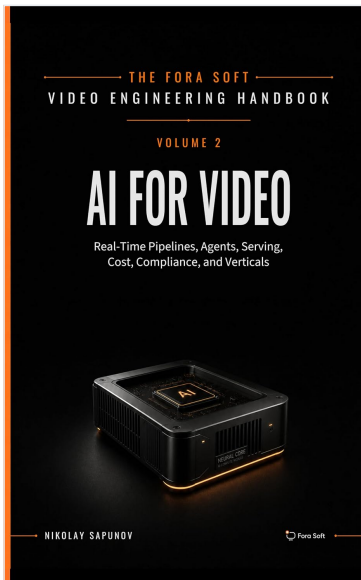
5 - Scoping checklist (ask these before you build, buy, or wrap)

- ☐ Is there a cheap-filter + sampler funnel, so the expensive reader sees ~tens of frames, not every frame?
- ☐ Are the heavy model calls routed through a batch API (50% off) because nothing is waiting?
- ☐ Does the job checkpoint progress, so a crash resumes at the next item instead of restarting?
- ☐ Does every work unit carry an idempotency key (video + model + policy version) to stop double-processing?
- ☐ Do failures retry, with a dead-letter queue isolating poison items for human inspection?
- ☐ Are verdicts versioned, so re-processing after a model/policy change is partial, not a full re-run?
- ☐ Do only low-confidence and high-severity items go to humans, with a random sample of confident ones spot-checked?
- ☐ Is the reason for each verdict stored (evidence + policy clause), to satisfy moderation-transparency law?

Companion to: 'Async Archive-Review Agent - Scoping & Durability Checklist'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/async-video-review-agent-archive-pattern

Fora Soft · www.forasoft.com



GO DEEPER · VOLUME 7 OF 7

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

From Real-Time Pipelines and Agents to Serving, Cost, Compliance, and Verticals

By Nikolay Sapunov, CEO at Fora Soft · 1,221 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0GX3DGBH4

This checklist scopes an agent that reviews a whole archive: the cost funnel, batch economics and a crash-proof durability spine. Volume 7 treats agents as engineering, with loop primitives, frameworks, memory and tools, plus serving on vLLM and Triton, cost per stream-hour and moderation duties.

WHAT YOU WILL FIND INSIDE

- Agent loop primitives, memory and tools
- Serving on vLLM, Triton and TensorRT
- Moderation and EU AI Act obligations
- Frameworks and orchestration for agents
- Cost per hour of processed video

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 an agent to review or tag a large video archive?

We will design the cost funnel and durable job pipeline for your archive and estimate cost per thousand hours before you build. 250+ video products shipped since 2005.

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