

AI Video Call — Architecture Blueprint

Fora Soft Learn · AI for Video Engineering

1 — The three planes: keep these jobs separate

Media plane	Carries audio + video. WebRTC + an SFU. Tightest time budget.
AI plane	Every ML model. Lives in one of three homes (see below).
Control plane	Rooms, participants, signaling, who-speaks. Loosest budget.

2 — Where each AI feature runs (match the signal it needs)

Home	Features	How it works
ON-DEVICE (sender)	Background blur, noise suppression, AR/beauty filters.	Needs raw frames + raw audio. \$0 per participant. Best privacy.
CLOUD AI AGENT (joins as part.)	Live captions, translation, meeting notes, avatar lip-sync.	Subscribes to streams, publishes results back. Scales per minute.
SFU-SIDE (fan-out)	Content moderation only. Central, cannot be bypassed by a client.	Copy the stream to a worker — never make the SFU decode inline.

3 — Two budgets that run in parallel (with the arithmetic)

INTERACTIVE MEDIA	one-way target ~200 ms (whether people talk over each other) Capture 10 + On-device AI 35 + Encode 10 + Network 120 + Playout 25 = 200 ms → if network costs 120 ms, only ~80 ms is left for everything else. Stay small.
AI VOICE LOOP	hear-to-answer target ~800 ms (whether the assistant feels quick) Turn-taking 200 + STT 100 + LLM first token 300 + TTS first chunk 150 + Net 50 = 800 ms. Stream every stage; use a learned turn detector, not a silence timer.

4 — Build order: each step ships a working product

Step 1	Plain call — WebRTC + SFU. Solid audio/video + reconnect. No AI yet.
Step 2	On-device cleanup — background blur + noise suppression. Highest value/effort.
Step 3	AI agent — live captions, then translation, then notes (reuse the transcript).
Step 4	Specialized — moderation, in-call avatars, domain assistants + AI Act notice.

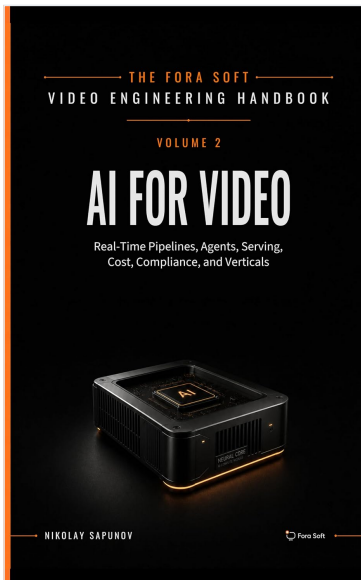
5 — Pre-launch checklist (engineering context, not legal advice)

- ☐ Exactly ONE noise suppressor in the chain — disable the browser built-in if you add yours.
- ☐ No audio AI after the Opus encoder; denoise on raw audio in an AudioWorklet, pre-encode.
- ☐ SFU never transcodes; moderation runs on a fan-out copy, not inline on the server.
- ☐ Written per-device latency budget; measure each on-device model on real mid-range phones.
- ☐ Disclose AI-generated or altered voice/face to participants — EU AI Act Article 50.
- ☐ Adopt the media plane and agent framework; build only the features + rules that are yours.

Companion to: 'Phase 6 Capstone — Building A Full AI-Enhanced Video Call'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/phase-6-capstone-full-ai-enhanced-video-call

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 blueprint separates the media, AI and control planes and gives the latency arithmetic for an AI video call. Volume 7 covers where AI plugs into a WebRTC pipeline, live captions and translation, agents in calls, serving and cost per stream-hour, and EU AI Act duties.

WHAT YOU WILL FIND INSIDE

- Where AI plugs into a WebRTC pipeline
- Agents that join calls as participants
- Live captions and translation in calls
- Cost per stream-hour and serving

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

Building a video call with live captions, translation or an AI agent?

We will lay out the three planes for your product, set the latency budgets and plan the build order. 250+ video products shipped since 2005.

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