Real-Time AI Conferencing — Build Blueprint

AI for Video Engineering · Capstone 10.1

1 · The reference stack: one SFU + the agent pattern

Media plane	Clients → SFU media server → clients. WebRTC + Opus. Forwards, never transcodes.
AI plane	One agent joins each room as a participant: STT → captions, translation, notes.
Edges + control	Token server (auth) · TURN/coturn (hard NAT) · app backend (transcripts).

2 · Build vs buy (2026): adopt infra, buy models, build the product

ADOPT	SFU (LiveKit / mediasoup) · agent framework (LiveKit Agents) · TURN (coturn)
BUY	speech-to-text (Deepgram / AssemblyAI) · translation (frontier LLM / MT API)
BUILD	on-device blur (MediaPipe + WebGPU) · token server · app backend · the UX
NOTE	Krisp went paid 2026-05-01; RNNoise unmaintained since 2024; DeepFilterNet free.

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

INTERACTIVE MEDIA	one-way ~200 ms. If network = 120 ms, only ~80 ms is left for capture + on-device AI + encode + playout. Keep on-device models small.
AI VOICE LOOP	~800 ms. VAD 50 + STT 150 + LLM 400 + TTS 150 + net 50 = 800 ms. Stream every stage; use a learned turn detector, not a silence timer.

4 · Scaling tiers and the cost rule

<100	→ single SFU node.
100–1K	→ simulcast + SVC (500–800 video/node).
1K+	→ cascaded SFUs across regions (a relay link is just another participant).
COST	on-device = $0 to operator · SFU media ~$0.0005/participant-min · agent ~$0.01/min. 25-person hour, all 5 features: ~$0 + ~$0.75 + ~$1.20 = under $2.00 (managed). Self-host the SFU past ~500 sustained concurrent sessions.

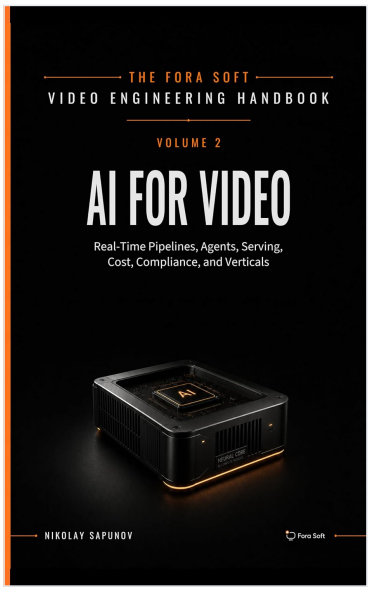
5 · Build order: each milestone ships a working product

1. Plain call (SFU+token+TURN)
2. On-device blur+denoise
3. Caption agent (streaming STT)
4. Translation + notes (same agent, same transcript)
5. Moderation/avatars + AI Act notice.

6 · Go-live checklist (engineering context, not legal advice)

- [ ] Plain call is solid first: audio/video + reconnect on the devices your users actually have.
- [ ] Per-person AI (blur, denoise) runs on the device — never in the cloud per participant.
- [ ] Exactly ONE noise suppressor on raw audio, before Opus; disable the browser built-in if you add yours.
- [ ] SFU never transcodes; moderation runs on a fan-out copy, not inline on the media server.
- [ ] TURN relay deployed with time-limited credentials; ~15–20% of networks need it to connect.
- [ ] Per-session observability (connect time, packet loss, agent latency) wired in from milestone 1.
- [ ] Disclose AI-generated or altered voice/face to participants — EU AI Act Article 50 (from 2026-08-02).

Companion to: 'Capstone — Real-Time AI Video Conferencing Platform'
www.forasoft.com/learn/ai-for-video-engineering/articles-ai/realtime-ai-conferencing-platform



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 lays out the SFU-plus-agent stack, build vs buy, two latency budgets, scaling tiers and a go-live checklist. Volume 7 covers this stack in depth: AI inside the WebRTC pipeline, live captions and translation, agents and tools, serving and cost per stream-hour, and the EU AI Act.

WHAT YOU WILL FIND INSIDE

- AI agents inside WebRTC rooms
- Live captions, translation and notes
- Serving and cost per stream-hour
- EU AI Act duties for voice and 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 a video conferencing product with AI built in?

We will ship it milestone by milestone: a solid call first, then on-device effects, captions, translation and moderation. 250+ video products shipped since 2005.

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