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)

```
ON-DEVICE (sender)  Background blur, noise suppression, AR/beauty filters.
                    Needs raw frames + raw audio. $0 per participant. Best privacy.
CLOUD AI AGENT      Live captions, translation, meeting notes, avatar lip-sync.
  (joins as part.)  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.