Real-Time Avatar Launch Checklist

Fora Soft Learn - AI for Video Engineering

1 - The loop: five stages racing one clock

```
Mic -> ASR -> LLM -> TTS -> Avatar synth -> WebRTC -> viewer. Stages are sequential.
Stream through them (start LLM/TTS/avatar early); do not wait for each to finish.
```

2 - The one number: turn-taking latency

```
Target the PERCEIVED gap near 200 ms; keep it under ~1.5 s or it feels robotic.
Budget to first frame ~= 150 ASR + 300 LLM + 150 TTS + 200 avatar + 100 WebRTC ~= 900 ms.
```

3 - Architecture: the avatar joins the call

```
Send only the agent's AUDIO to the avatar; let the avatar publish video into the room.
Avoids a round trip and double video encoding (the naive WebSocket path pays it twice).
Handle interruptions + playback-done over an RPC/data channel, in milliseconds.
```

4 - Buy vs build (2026 - re-verify before you commit)

```
Buy (fastest to ship): Tavus Phoenix-4 (sub-600 ms), HeyGen LiveAvatar (1-2 s), Simli, bitHuman.
Self-host (face control / data residency / high volume): MuseTalk + LivePortrait (MIT), NVIDIA ACE.
Decide on latency need, control over the face, and per-minute volume vs a GPU lease.
```

5 - Disclosure: settle this BEFORE you ship in the EU

```
EU AI Act Article 50 treats a synthetic face as a deepfake: disclose + mark machine-readable.
Transparency obligations apply from 2 August 2026. Live video: persistent 'AI' label + opening note.
```

6 - Pre-launch checklist

- [ ] Measured the real perceived turn-taking gap end to end (not just one stage's latency).
- [ ] Avatar renderer joins the call as a participant; agent sends audio only, no video round trip.
- [ ] Interruptions cancel in-flight frames and switch to a listening pose within ~200 ms.
- [ ] Picked buy vs build against latency need, face control, and streamed-minute volume.
- [ ] Locked the voice path (streaming TTS or speech-to-speech) and tested it under interruption.
- [ ] Added the EU AI Act Article 50 disclosure: persistent 'AI' label + opening disclosure.
- [ ] Confirmed consent for any cloned face or voice before going live.