AI Deployment Topology Cheat Sheet

Latency budget × Deployment topology × Real-time vs batch — for every video-AI feature.

1. Latency budgets by user-experience class

Live captions (partial)	300 ms
Live captions (final)	1.0 s
Live translation	800 ms
AI agent in call	200–500 ms
Background blur per frame	33 ms
In-call action items	3–10 s
Meeting summary	30 s – 5 min
Highlight reel	1–10 min
Archive search index	hours / nightly

2. Topology layers — round-trip + cost-per-minute envelope

LAYER	ROUND-TRIP	COST / MIN / STREAM	TYPICAL FEATURE
On-Device	0 ms	$0 / min	Background blur, on-device blur, noise
Edge	5–30 ms	$0.0003–0.003 / min	Live captions, translation, in-call agent
In-Region Cloud	30–100 ms	$0.001–0.05 / min	Meeting summary, generative video
Cross-Region Cloud	100–300 ms	$0.001–0.05 / min	Regulatory or model-availability only

3. Five-step decision tree

1. What is the user-experience latency budget?
2. Max tolerable round-trip = budget − inference window.
3. Which topology layer fits the model size?
4. Cost-per-minute envelope at expected concurrency?
5. Streaming partial answers, or batched single answer?

4. Three recurring mistakes

On-device-as-cloud: blur in cloud GPU when on-device costs $0.
Live-summary: summarising mid-call when batched post-call is better.
Cross-region default: us-east-1 for EU users — fatal for real-time.