

Automotive — Cross-border ADAS Serving

Client: Tier-1 European OEM · **Region:** EU / CN

Workload

Vision transformer (1.8B) + trajectory model, 42k QPS peak

Deployment topology

6 Sovereign Edge PoPs · 2 Core IDC (Frankfurt, Shanghai)

Measured results

Metric	Value	Note
p95 latency	38 ms	from 112 ms
Inference cost	-47%	per 1M tokens-equiv
Data residency	100%	in-region

Methodology

Figures anonymised at customer request. Latencies measured at the client SDK boundary over rolling 14-day windows. Cost deltas reference the customer's prior production stack, normalised per token-equivalent.