

Архитектурный worksheet anomaly detection

Пройдите пятислойный стек против четырёх чисел — латентность, камеры, метки, режим.

1. Четыре числа — заполните в первую очередь

Бюджет латентности

Sub-150 мс → edge.
200–500 мс → гибрид.
>500 мс → cloud.

ваш ответ

Число камер

<20 → SaaS вероятно.
20–100 → гибрид.
100+ → custom вероятно.

ваш ответ

Размеченные примеры

<100 → CLIP zero-shot.
100–10K → MIL.
>10K → fully supervised.

ваш ответ

Регуляторный режим

EU AI Act → edge-first.
GDPR ст. 9 / ст. 35.
HIPAA / BIPA / CCPA.

ваш ответ

2. Пятислойный стек — отметьте слои, нужные вашей сборке

Слой	Где	Типовая модель	Латентно сть	Когда пропустить
1. Candidate filter	Edge	I3D / SlowFast INT8, MemAE	30–80 мс	Никогда
2. Class classifier	Edge или cloud	MIL: RTFM, MGFN, MTFI	30–60 мс	Чистая object-detection задача
3. VLM rescorer	Cloud	VadCLIP, TPWNG, VadCLIP++	100–300 мс	Один фикс. класс и много меток
4. Frontier explanation	Cloud	Gemini 2.5 Pro, Claude Opus 4.7	1–3 с	Нет регулятора или audit
5. Signal anomaly	Sidecar	Isolation Forest, LSTM-AE	Сек – мин	Никогда — он ловит тихие отказы

3. Класс аномалии — назовите до выбора модели

- ☐ Поведенческая — бег, лазание, падение. 3D-CNN + MIL.
- ☐ Физическая — брошенная сумка, лужа, лишний объект. YOLO + tracking.
- ☐ Плотность / толпа — переполненная очередь, давка. Heat-map регрессия.
- ☐ Расписание / время — движение в 03:00, открытая дверь. Правила + норма.
- ☐ Статистическая — замёрзший поток, дрейф энкодера. Isolation Forest по метрикам.
- ☐ Комбинация — несколько из перечисленных. Слоистый стек.

4. Решение по топологии

Только edge

Jetson Orin Nano Super или Hailo-8 на камеру. \$200–600 за устройство. Sub-200 мс. Кадры не покидают площадку. Сильнейший AI Act posture.
Подходит: маленькая площадка, регулируемая, плохая связность.

Гибрид (дефолт)

Edge гонит слои 1–2 на high-precision; cloud делает rescore top 5–20%. Сокращение cloud-объёма в 6 раз, удержание 95%+ настоящих аномалий.
Подходит: 20+ камер, real-time SLA, регулируемые кадры.

Только cloud

Стрим всех камер на L4 / L40S / A100 в managed-регион. \$1–2.50 за GPU-час. 100+ потоков на GPU.
Подходит: forensic-only, центральное переобучение, нет sub-300 мс SLA.

5. EU AI Act + audit trail — инженерные артефакты, не бумаги юристов

High-risk под Annex III (статьи 9–15) → обязательно: документированный риск-менеджмент, dataset governance, тех. документация, автоматическое логирование, human oversight, требования к точности и устойчивости, conformity assessment. Штрафы до 3% глобального оборота или €15 млн. Инженерные артефакты: model card, data sheet, training log, drift log, alert audit trail, очередь ревью с sign-off, immutable storage (S3 Object Lock). Предпочтительный класс: gait, posture, object, поведение — избегайте face-based, где use case позволяет. Статья 50 — раскрытие AI-сгенерированного или AI-augmented контента.