

1 — Эталонный стек: каскад поиска + агент

Всегда вкл. Камеры -> VMS -> edge-детектор (YOLO11/Jetson) -> трекинг+аномалии -> индекс.
По запросу Агент планирует -> индекс даёт короткий список -> VLM читает -> пишет вывод.
Правило Дешёвый детектор по всему; дорогая модель читает лишь несколько клипов.

2 — Build vs Buy (2026): интегрировать инфру, брать/купить модели, строить агента

ИНТЕГРАЦИЯ VMS (Milestone / Genetec / Eagle Eye) — запись и хранение решены
ВЗЯТЬ детектор (YOLO11 на Jetson/DeepStream) · трекер (ByteTrack / SAM 2)
КУПИТЬ эмбединги+индекс (вектор-БД / Twelve Labs) · VLM-читатель (Qwen-VL)
СТРОИТЬ логику агента (LangGraph) · приложение · аудит · UI ревью = продукт
ВАЖНО стабильный YOLO11, HE v12/v13 (рекомендация Ultralytics, 2026).

3 — Арифметика каскада (почему это доступно)

1080p @ 15 fps = 15 x 3600 x 24 = ~1 300 000 кадров / камера / сутки.
Каждый кадр -> frontier-модель @ \$0.01/кадр = ~\$13 000 / камера / сутки (абсурд).
Каскад: детектор по всем -> ~1500 событий -> индекс -> ~20 клипов/запрос -> VLM.
Итог: модель видит ~20 клипов, а не 1,3 млн кадров. Запрос стоит сильно меньше \$1.

4 — Модель затрат: 50 камер, 500 расследований / месяц

EDGE ~4x Jetson Orin AGX (8-16 потоков). Железо — по числу камер, не нагрузке.
ИНДЕКС По активности, не длине видео. Нижние сотни \$ / месяц для площадки.
РАССУЖДЕНИЕ ~\$0.50-1.00 / расследование. 500 x \$0.75 = ~\$375 / месяц времени VLM.
ФОРМА С использованием растёт лишь рассуждение; edge фиксирован.

5 — Порядок сборки: каждый майлстоун выкатывает рабочий продукт

- 1 Поиск по записи (VMS + детектор + трекер + индекс + строка поиска).
- 2 Добавить VLM-читателя: топ-клипы -> фраза. 3 Цикл агента: многошаговые расследования.
- 4 Слой ревью + аудита (человек в цикле, дело, цепочка хранения).
- 5 Регулируемые расширения (аномалии, межкамерный, личность) — сперва против закона.

6 — Чек-лист запуска и закона (инженерный контекст, не юр. консультация)

- ☐ Дешёвый детектор по каждому кадру; VLM читает лишь короткий список — не каждый кадр.
- ☐ Эмбединги вычислены при обнаружении; поиск — мгновенный, а не пере-скан.
- ☐ Человек подтверждает каждый вывод; VLM уверены, даже когда ошибаются; журнал аудита и шифрование.
- ☐ EU AI Act: живая идентификация лиц в публичном месте = ЗАПРЕЩЕНО (Ст. 5, с 2025-02-02).
- ☐ Пост-фактум идентификация лиц = высокий риск (Прил. III, с 2026-08-02); поиск по описанию — лёгкая коробка.