# Чек-лист обслуживания инференса видео-AI

Fora Soft Learn · AI для видео-инженерии



## 1 — Две идеи, которые ускоряют LLM/VLM-сервер

```
PagedAttention      KV-cache маленькими страницами по запросу -> потери с 60-80% до менее 4%.
Continuous batching новый запрос встаёт сразу, как освободится место -> GPU с 30-40% до 75-90%.
Вместе: ~2-4x throughput прошлого поколения серверов, на ТОЙ ЖЕ карте.
```

## 2 — Варианты обслуживания (open + NVIDIA + serverless)

```
vLLM            дефолт для LLM и VLM; много пользователей; OpenAI-совместим; Docker; NVIDIA/AMD/TPU.
Ollama          один пользователь, одна машина (ноутбук / Apple Silicon). Прототип — НЕ для парка.
SGLang          много общих префиксов (рубрика, RAG); ~29% над vLLM, до 6.4x на префиксах.
TensorRT-LLM    макс. скорость на NVIDIA (15-30% над vLLM). 2026: PyTorch-бэкенд, без компиляции engine.
Triton          много моделей на парке; ensembles = конвейеры в сервере. Преемник: NVIDIA Dynamo (2025).
Serverless      Modal / RunPod / Replicate; масштаб до нуля; оплата по секундам; cold start на больших.
```

## 3 — Видео-функция — это конвейер: каждому этапу свой сервер

```
Декодирование (кадры+аудио)   FFmpeg / декодер.
Детекция / сегментация        CV-модели (YOLO, SAM) -> TensorRT, фикс. вход -> фикс. выход, 1 проход.
Транскрипция (речь-в-текст)   faster-whisper на CTranslate2 (~4x быстрее; ещё быстрее батчем).
Описание (vision-language)    VLM -> vLLM (или SGLang / TensorRT-LLM).
Резюме (язык)                 LLM -> vLLM. Свяжите всё для парка — ensemble в Triton.
```

## 4 — Реальность затрат (один пример батчинга)

```
85% занят / 35% занят = 2.4x больше токенов с той же карты.
$3.00/ч / 1.0M ток = $3.00 за млн;  $3.00/ч / 2.4M ток = $1.25 за млн  ->  на 58% дешевле, та же GPU.
```

## 5 — Перед утверждением плана обслуживания спросите:

- [ ] Каждый этап обслуживается софтом под свой тип модели, а не одним сервером под один этап?
- [ ] Языковые/vision-language этапы на vLLM (или SGLang/TensorRT-LLM), с PagedAttention + continuous batching?
- [ ] Ollama ограничен прототипированием и НЕ стал продакшен-архитектурой по инерции?
- [ ] Речь обслуживается специализированным движком (faster-whisper/CTranslate2), а не LLM-сервером?
- [ ] CV-модели скомпилированы в TensorRT для реального времени, инференс за один проход?
- [ ] Если много моделей на парке — они связаны как ensemble в Triton (или через Dynamo на масштабе ЦОД)?
- [ ] Всплесковый/малый трафик на serverless (масштаб до нуля), стабильный/большой — на «тёплом» кластере?
- [ ] Считаем стоимость ЗА ТОКЕН, а не только задержку — и измерили выигрыш батчинга на своей нагрузке?