# Памятка по бюджету кадров видео-VLM

Fora Soft Learn · AI for Video Engineering



## 1 · Стоимость — это два умножения

Видео-VLM никогда не видит ваше видео — только сэмплированные кадры в виде токенов. Стоимость = (кадры × токены-на-кадр + звук) ÷ 1 000 000 × цена-за-миллион.

```
10-мин видео @ 1 fps, Gemini, разрешение по умолчанию:
  600 кадров  x 258 токенов/кадр = 154 800 визуальных токенов
  600 секунд  x  32 токена/сек   =  19 200 аудиотокенов  -> ~174 000 входных токенов
  174 000 / 1 000 000 x $0,30/M  = ~ $0,052 за видео
Низкое разрешение (66 токенов/кадр): ~59 000 токенов -> ~ $0,018 за видео
```

## 2 · Контекстное окно — жёсткая стена

Модель на 1 миллион токенов вмещает ~1 час видео при разрешении по умолчанию или ~3 часа при низком (доки Google Gemini, 2026). Сверх — снизьте разрешение, снизьте частоту кадров или переходите на поток.

## 3 · Сэмплирование кадров vs поток токенов

### Сэмплирование (офлайн)

- Всё видео разом; рассуждает по нему целиком.
- Ограничено контекстным окном (~1-3 ч).
- Платите один раз за видео, авансом.
- Лучше: записи встреч, лекции, обзор архива.

### Поток токенов (онлайн)

- Кадры подаются непрерывно; старые угасают.
- Тянет бесконечное видео; помнит лишь недавнее.
- Платите непрерывно, пока идёт поток.
- Лучше: живые камеры, звонки, оповещения.

## 4 · Чек-лист перед тем, как строить

- [ ] Насколько коротко самое короткое событие, которое функция обязана поймать? Меньше секунды — 1 fps пропустит, поднимите частоту.
- [ ] Ответ нужен вживую или потом? Вживую = поток токенов; потом = сэмплирование кадров.
- [ ] Влезет ли всё видео в контекстное окно при вашей частоте и разрешении? Если нет — режьте на куски или стримьте.
- [ ] Дефолтное или низкое разрешение? Низкое режет визуальные токены в ~4x и годится для медленного контента (лекции, статичные камеры).
- [ ] Для длинной неотобранной записи нужно ли адаптивное сэмплирование? Оно добавляет ~8-10 пунктов точности над равномерным 1 fps.