

1 - Сначала выберите лагерь, потом модель

- ☐ Контент обязан оставаться на наших серверах (медицина, юр., непубличный материал)? Если да -> self-host open weights.
- ☐ Объём высокий и устойчивый, так что фикс-цена GPU выгоднее аренды за клип? Если да -> open weights выигрывают за точкой перелома.
- ☐ Если оба ответа «нет» (так у большинства продуктов на старте) -> арендуйте closed API и переходите к продукту.

2 - Сначала отсев по жёстким ограничениям, качество — потом

- ☐ Отсев ПО: размещение данных, макс. латентность, макс. цена за клип, нужное разрешение и длина клипа.
- ☐ Только ПОТОМ ранжируйте оставшихся по качеству (Elo Artificial Analysis Video Arena). Популярная, лучшая и самая дешёвая — обычно 3 разные модели.
- ☐ Под задачу: монтаж готового видео -> Runway (Aleph V2V). Про-колор / HDR -> Luma Ray3. Уже на Google Cloud -> Veo 3.1. Топ слепых тестов -> Kling 3.0. Лучшая цена/качество -> Hailuo 02. Много черновиков -> Pika 2.5. Бренд + связные диалоги -> Sora 2.

3 - Стройте async pipeline сразу, в первой версии

Pipeline: Submit -> [МОДЕРАЦИЯ промпта] -> Очередь (job ID) -> Генерация (poll 10-20с ИЛИ webhook)
-> [МОДЕРАЦИЯ результата] -> Метка (C2PA + watermark) -> Хранение + выдача.

Барьер 1: модерируйте промпт ДО отправки — после submit вы уже оплатили задачу.

Барьер 2: сохраняйте метки происхождения через любой перекодек/правку — не срывайте их.

Тонкий внутренний слой перед вендором: failover / замена устаревшей модели без переписывания.

4 - Посчитайте стоимость один раз и поставьте потолок

- ☐ За клип = сек x \$/сек. В месяц = клипы x за клип. Реально в месяц = x (генераций на сохранённый клип) — обычно ~3.
- ☐ Пример: 8с x \$0.10/сек x 10 000 клипов = \$8 000/мес -> x3 ретрая = \$24 000/мес. Планируйте по числу x3.
- ☐ Главные рычаги: тариф модели (бюджет \$0.03/сек срезает ~2/3) и число ретраев (дешёвый черновик, премиум только на финал).
- ☐ Поставьте жёсткий потолок расходов и лимит циклов на задачу, чтобы сбойная фича не съела бюджет запуска.

5 - Комплаенс и лицензии — закройте до запуска

- ☐ EU AI Act статья 50 (с 2 авг. 2026): помечать + раскрывать AI-видео. C2PA — принятый способ маркировки.
- ☐ Closed-модели (Sora 2, Veo) уже подписывают результат по умолчанию — ваша задача СОХРАНИТЬ метку до зрителя.
- ☐ Лицензии open weights различаются: Apache 2.0 (Wan, LTX-Video, CogVideoX) = полная коммерция; HunyuanVideo исключает EU/UK и гейтит крупных деплоеров. Читайте до сборки.