

1 — Налог на шаги (зачем это нужно)

стоимость клипа ≈ стоимость шага × число шагов
50 шагов -> ~50 единиц GPU 4 шага -> ~4 единицы => ~12x меньше работы, цены и ожидания

2 — Одна идея: дистилляция

- ☐ Медленная точная модель-УЧИТЕЛЬ обучает быстрого few-step УЧЕНИКА прыгать к тому же ответу.
- ☐ LoRA — небольшая накладываемая заплатка (десятки МБ), которую цепляют на базовую модель без переобучения.
- ☐ CFG (classifier-free guidance) — регулятор давления промпта; дистиллированные модели запекают его внутрь — выключите (обычно 1.0).

3 — Методы (сначала читайте «для чего»)

Метод	Шагов	Для чего
LCM / LCM-LoRA	1-4	Быстрая накладка на готовую модель изображений
SDXL-Turbo (ADD)	1-4	Доказательство одношаговой скорости
SDXL-Lightning	1-8	Настраиваемый баланс скорости/качество на SDXL
Hyper-SD	1-8	Лучшее одношаговое качество изображений
AnimateDiff-Lightning	1-8	Быстрое короткое видео во многих стилях
Wan2.2-Lightning	4-8	Ускорить open-видеомодель, которую вы запускаете
CausVid / Self Forcing	few	Видео в реальном времени, потоковое

4 — Бюджет шагов из опыта, затем метод

- ☐ Батч / ночью (никто не ждёт): полные шаги (25-50), без ускорения — качество бесплатно.
- ☐ Интерактив (пользователь жмёт, смотрит спиннер): 4-8 шагов через дистиллированную модель. Изображение -> LCM-LoRA, или Hyper-SD для лучшего 1-шага. Видео -> AnimateDiff-Lightning, или Lightning-дистилляция вашей модели.
- ☐ Живой / реальное время (реагирует на движения): causal-поток — CausVid / Self Forcing, или нативная модель реального времени LTX-2.

5 — Подводные камни перед сборкой

- ☐ Выцветшие дистиллированные клипы почти всегда значат, что CFG оставили включённым — ставьте значение дистиллированной модели (обычно 1.0).
- ☐ Меньше шагов — меньше тонких деталей и богатства движения; 4 шага это частая продакшен-середина, 1 шаг — самый резкий размен.
- ☐ Лицензии меняются: дистиллированная LoRA или версия модели может нести иную лицензию, чем база — проверяйте ИМЕННО ту версию, что поставляете.