Vision Transformer Decision Worksheet

Четыре вопроса, три ручки, четыре ошибки — выбрать ViT-backbone за 90 секунд.

Дерево из четырёх вопросов

1. Какая задача?

- ☐ Классификация → ViT-S/B с DINOv3 pretrain.
- ☐ Сегментация → SAM 2 (Hiera backbone).
- ☐ Детекция → RT-DETR или Grounding DINO.
- ☐ Видео-классификация → Hiera-MAE или VideoMAE-V2.
- ☐ Vision-language → выбирайте VLM, не backbone.

2. Бюджет задержки?

- ☐ < 50 мс edge → ViT-Tiny/Small при 224.
- ☐ 50–200 мс облако → ViT-Base/Large при 224 или 384.
- ☐ 200 мс+ batch → ViT-Huge при 384 или больше.

3. Объём размеченных данных?

- ☐ < 10 K → frozen backbone, fine-tune только голова.
- ☐ 10 K–100 K → fine-tune последние слои.
- ☐ > 100 K → полный fine-tune, низкий LR.
- ☐ Никогда не обучайте ViT с нуля на малых данных.

4. Требования к лицензии?

- ☐ Коммерч. → Hiera, DINOv3, EVA-02, SigLIP 2 (MIT/Apache).
- ☐ Non-commercial? → TimeSformer ок, остальные — review.

Три ручки, решающие throughput

Ручка	Диапазон	Эффект
Patch size	14, 16, 32	Меньше = больше токенов
Разрешение	224, 384, 512+	Стоимость квадратично
Embed dim	192–1408	Ёмкость vs throughput
Размер модели	T / S / B / L / H	T,S edge; L,H облако

Четыре ошибки — чек-лист перед запуском

- ☐ Квадратичный attention протестирован на целевом разрешении.
- ☐ Position embeddings интерполированы при смене разрешения.
- ☐ Pretrained checkpoint — никогда не с нуля на малых данных.
- ☐ Edge-экспорт (ONNX, TensorRT, CoreML) проверен рано.

Reference checkpoints (Apache 2.0 / MIT)

- ☐ Dense-задачи на изображениях → DINOv3 (Meta, 2025).
- ☐ Сегментация изображений + видео → Hiera (Meta, ICML 2023).
- ☐ Видео-классификация → VideoMAE-V2 (CVPR 2023).
- ☐ Image-энкодер для VLM → SigLIP 2 (Google, 2025) или EVA-02.