Vision Transformer Decision Worksheet

Four questions, the three knobs that matter, and four pitfalls — pick a ViT backbone in ninety seconds.

Four-question decision tree

1. What is the task?

- ☐ Classification → ViT-S/B with DINOv3 pretrain.
- ☐ Segmentation → SAM 2 (Hiera backbone).
- ☐ Detection → RT-DETR or Grounding DINO.
- ☐ Video classification → Hiera-MAE or VideoMAE-V2.
- ☐ Vision-language → pick the VLM, not the backbone.

2. Latency budget?

- ☐ < 50 ms on edge → ViT-Tiny/Small at 224.
- ☐ 50–200 ms cloud → ViT-Base/Large at 224 or 384.
- ☐ 200 ms+ batch → ViT-Huge at 384 or higher.

3. Labelled data volume?

- ☐ < 10 K → freeze backbone, fine-tune head only.
- ☐ 10 K–100 K → fine-tune last few layers.
- ☐ > 100 K → full fine-tune, low learning rate.
- ☐ Never train ViT from scratch on small data.

4. Licence requirements?

- ☐ Commercial → Hiera, DINOv3, EVA-02, SigLIP 2 (MIT/Apache).
- ☐ Non-commercial only? → TimeSformer fine, others need review.

Three knobs that decide throughput

Knob	Range	Effect
Patch size	14, 16, 32	Smaller = more tokens, more cost
Resolution	224, 384, 512+	Cost scales quadratically
Embed dim	192–1408	Capacity vs throughput trade-off
Model size	T / S / B / L / H	T,S for edge; L,H for cloud

Four pitfalls — pre-launch checklist

- ☐ Quadratic attention tested at target resolution.
- ☐ Position embeddings interpolated if resolution differs.
- ☐ Pretrained checkpoint used — never trained from scratch on small data.
- ☐ Edge export path (ONNX, TensorRT, CoreML) validated early.

Reference checkpoints (Apache 2.0 / MIT)

- ☐ Image dense tasks → DINOv3 (Meta, 2025).
- ☐ Image + video segmentation → Hiera (Meta, ICML 2023).
- ☐ Video classification → VideoMAE-V2 (CVPR 2023).
- ☐ VLM image encoder → SigLIP 2 (Google, 2025) or EVA-02.