1. The DeepStream Hardware-Acceleration Map

In a production environment, you must ensure that every stage of the pipeline stays on the specific hardware silicon designed for it. This ensures the CPU is free for high-level logic.

Silicon Offloading Logic

- **Decoding:** Handled by **NVDEC**. Always set your decoder to use unified-memory to prevent the "data shuffle" between CPU and GPU.
- **Preprocessing:** Handled by the **VIC (Video Image Compositor)**. This is where you perform the 1080p to 512x512 scaling for your segmentation model without touching the GPU's CUDA cores.
- **Inference:** Handled by **TensorRT on GPU cores**. This is the heavy lifting for YOLOv4.
- **Depth & Stitching:** Handled by the **DLA (Deep Learning Accelerator)**. Since DLA is independent of the GPU, you can run your 3D reconstruction and stitching here in parallel with detection.
- **Rendering:** Handled by **EGL**. Use this for the on-device overlay so that the visualization doesn't lag the actual detection.

---

2. NVIDIA TAO (Transfer Learning) Command Set

To fine-tune your model for specific industrial safety gear (like a specific brand of high-visibility vest), use this sequence in your dev environment.

The Fine-Tuning Sequence

1. **Dataset Conversion:** Convert your annotated images (in KITTI format) into TFRecords for TAO efficiency.
    - tao model yolo_v4 dataset_convert -d specs/dataset_config.txt -o data/tfrecords/
2. **Model Training:** Load the pre-trained YOLOv4 weights and apply your industrial dataset.
    - tao model yolo_v4 train -e specs/train_config.txt -r results/experiment_1/ -k $API_KEY
3. **Model Pruning:** Remove the "weak" connections in the neural network to make it run faster on the Jetson Orin Nano.
    - tao model yolo_v4 prune -m results/experiment_1/weights.tlt -o results/pruned_model.tlt -eq union -pth 0.1
4. **INT8 Quantization:** This is the most important step for industrial speed. It converts the model to 8-bit precision.
    - tao model yolo_v4 export -m results/pruned_model.tlt -o results/yolov4_int8.etlt --data_type int8 --cal_cache_file results/cal.bin

---

3. Industrial Reliability Manifest

This document outlines the "Safety First" software rules for the plant floor.

Watchdog & Recovery Logic

- **System Integrity:** Implement a **systemd** service with Restart=always and a StartLimitIntervalSec=0. This ensures that if the DeepStream app crashes due to a camera signal loss, it attempts an immediate reboot.
- **Network Resilience:** Use a **Local-First MQTT strategy**. The Jetson should write all safety violation metadata to a local SQLite ring buffer (keeping only the last 24 hours). A background process should sync this to the cloud only when a "Heartbeat" to the server is confirmed.
- **Thermal Throttling Graceful Degradation:** Instead of letting the Orin shut down at 85°C, script a trigger that detects tegra_stats temperatures. If heat exceeds 80°C, the system should automatically switch to a "High Latency" mode (dropping from 30 FPS to 10 FPS) to shed thermal load while maintaining basic safety monitoring.