On-Device CV Deployment Checklist

Depth Anything V2 + SmolVLM2 — pick the right variants and avoid the licensing and pipeline traps.

1. Pick the right Depth Anything V2 variant

- [ ] Commercial product? -> Depth-Anything-V2-Small ONLY (Apache 2.0, 24.8M params, 50+ FPS on consumer GPU).
- [ ] Research / internal eval only? -> Base / Large / Giant available, but they are CC-BY-NC-4.0 (no commercial use).
- [ ] Indoor scenes (offices, telemedicine, e-learning)? -> Metric-Hypersim variant, 20 m max depth.
- [ ] Outdoor scenes (surveillance, drone, automotive)? -> Metric-VKITTI variant, 80 m max depth.
- [ ] Video, stored clip needs consistent depth? -> Use Video Depth Anything (CVPR 2025) instead of frame-by-frame V2.
- [ ] Live stream needs low latency at 2K? -> Use FlashDepth (2025) — V2 backbone + recurrent temporal alignment.

2. Pick the right SmolVLM2 size

- [ ] Browser tab / WebGPU? -> SmolVLM2-256M, <1 GB GPU RAM, ~80 decode tokens/sec on M4 Max.
- [ ] Smartphone (iPhone 13+, Pixel 8+)? -> SmolVLM2-500M, on-device video analysis, Hugging Face demo app available.
- [ ] Jetson Orin / laptop / consumer GPU? -> SmolVLM2-2.2B, ~5.2 GB GPU RAM, best video understanding in class.
- [ ] License? -> All three sizes are Apache 2.0. No NC carve-out. Commercial use is fine.

3. Engineer the pipeline as a funnel

- [ ] Stage 1 — Motion gate (background subtraction / tiny YOLO-N) drops ~99% of frames on the camera NPU.
- [ ] Stage 2 — Depth + detection in parallel; combine to get a 3D position; gate by geometric zone.
- [ ] Stage 3 — SmolVLM2 with a structured yes/no prompt, only on the rare frames that earn the cost.
- [ ] Pair every SmolVLM call with a second pass on a different sampled frame; only act when the two agree.
- [ ] For high-stakes decisions, route survivors to a frontier VLM on the server; on-device VLM is the cheap filter.

4. Six failure modes to audit before going live

- [ ] Flicker in depth track -> Vanilla V2 on video wobbles; switch to Video Depth Anything or FlashDepth.
- [ ] Wrong metric variant -> Hypersim indoors, VKITTI outdoors; route at runtime if the camera covers both.
- [ ] License trap -> Prototype with Large, ship with Small. Audit every checkpoint shipped in the binary.
- [ ] VLM hallucination -> Constrain output with structured prompts; require agreement across sampled frames.
- [ ] Asking VLM for geometry -> Never ask a small VLM for absolute distances; use depth model for numbers.
- [ ] Memory floor -> Need 4 GB device RAM + AI accelerator; iPhone 13 / Pixel 8 / Jetson Orin Nano clear it.