

Edge Compression - Distillation + Quantization Checklist

Fora Soft Learn · AI for Video Engineering

1 - The two levers that make a model smaller

Quantization	keep the model, store its numbers in fewer bits (FP32/16 -> INT8/INT4). Cheap to apply.
Distillation	train a NEW smaller 'student' model to copy a bigger 'teacher'. Powerful but costly to create.

Use BOTH: distil first, then quantize the student -> the footprint + speed savings multiply.

2 - The memory math (a 7-billion-parameter model)

FP32 (4 B/wt) = 28 GB FP16 (2 B/wt) = 14 GB INT8 (1 B/wt) = 7 GB INT4 (0.5 B/wt) = 3.5 GB
 8 GB edge ceiling (Jetson Orin Nano): FP16 won't load; INT8 just fits; INT4 fits with headroom.
 INT8 ~4x smaller than FP32, usually under 1% accuracy lost. INT4 needs AWQ/GGUF to stay accurate.

3 - The quantization methods (recognise these on a spec sheet)

INT8 + calibration	8-bit; edge via TensorRT; under 1% loss. The safe default for computer-vision models.
AWQ	4-bit; protects ~1% salient weight channels; near-FP16 quality; great for VLMs/LLMs.
GPTQ	4-bit; layer-by-layer error correction; the other mature 4-bit route.
GGUF Q4_K_M	~4.5-bit; llama.cpp on laptop/Mac/box; ~2% quality cost for ~40% memory saved.
FP8 / NVFP4	8-/4-bit float on recent NVIDIA; NVFP4 on Blackwell: <=1% loss, up to 4x FP8 speed.
PTQ vs QAT	PTQ first (minutes, no retrain). QAT only if PTQ's accuracy loss is too large to ship.

4 - A video feature is a pipeline - compress each stage

Detect / segment (YOLO, SAM)	INT8 via TensorRT - PTQ first, QAT if accuracy on rare classes slips.
Transcribe (speech-to-text)	a distilled model (Distil-Whisper), then quantized to INT8.
Describe / summarize (VLM/LLM)	4-bit weights via AWQ or GGUF - the largest models, the biggest win.

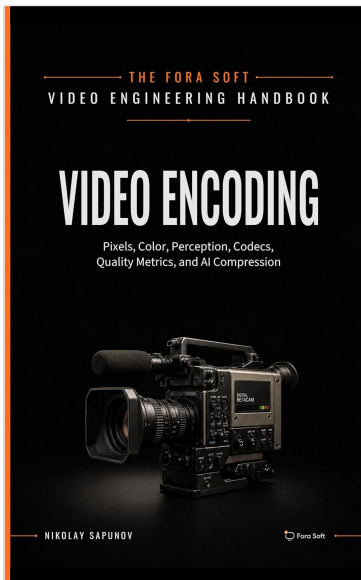
5 - Before you compress, ask:

- ☐ Does this feature need to run on the edge at all (volume, bandwidth, latency, privacy) - or is the cloud fine?
- ☐ Have we tried the cheap lever first: post-training INT8 quantization, before any distillation or QAT?
- ☐ Are the big language / vision-language stages on 4-bit (AWQ or GGUF), where the memory math bites hardest?
- ☐ For speech, are we using a distilled model (Distil-Whisper) rather than the full teacher - then quantizing it?
- ☐ Did we re-measure accuracy on OUR footage after every compression step, not trust a public-benchmark average?
- ☐ Did we check the rare-but-important classes (the weapon, the fall, the defect), not just average accuracy?
- ☐ Does the model fit the device's memory ceiling (often 8 GB) with room left for its working memory?
- ☐ Is the model exported to the right runtime for the chip (TensorRT/INT8 on Jetson; Core ML / NNAPI on phones)?

Companion to: 'Edge Compression - Distillation + Quantization Checklist'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/distillation-quantization-for-edge-video-ai

Fora Soft · www.forasoft.com



GO DEEPER · VOLUME 1 OF 7

Video Encoding: A Complete Guide to Video Compression

From Pixels and Color to H.264, HEVC, AV1, Quality Metrics, and AI

By Nikolay Sapunov, CEO at Fora Soft · 1,522 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HF52B65R

This checklist decides how to shrink a vision or speech model for the edge with distillation and quantization. Volume 1 covers the video side of the same edge pipeline: compression from pixels and color to H.264, HEVC and AV1, quality metrics and AI in video coding.

WHAT YOU WILL FIND INSIDE

- H.264, HEVC and AV1 encoding
- Quality metrics for compressed video
- AI in video compression
- Pixels, color and bit depth

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Your model does not fit on the Jetson or the phone?

We will quantize or distill it, re-measure accuracy on your footage and export it to the right runtime for the chip. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)