AI inside the encoder - cheat sheet

Fora Soft Learn

Four ML drop-ins, the encoders that ship them, BD-rate numbers - one page

Companion to article 6.3 of Block 6. Print on A4.

1. The four levers

Lever	Lives in	Model	Headline gain
Scene-cut	Lookahead	Small classifier	1-3% BD-rate
Partition gate	Partition stage	LightGBM or CNN	30-82% time saved
Mode-decision	Mode decision	Ranking classifier	40-70% time saved
Perceptual QP	Quantize	Variance, saliency	+3 to +8 VMAF

2. Open-source encoder support in 2026

Encoder	Partition ML	Mode ML	Scene-cut	Variance-AQ	Psy-RD
x264	no	no	classical	yes	yes
x265	no	limited	classical	yes	yes (default 2.0)
SVT-AV1	yes	yes	classical	yes	psy fork only
libaom	limited	no	classical	yes	no
VVenC	yes	limited	classical	yes	no

3. FFmpeg flags that activate each lever

Lever	x264 flag	x265 flag	SVT-AV1 flag
Scene-cut	--scenecut 40 (default)	--scenecut 40	auto, --keyint cap
Partition gate	via --preset slow	via --preset slower	via -preset 4-8
Mode ranking	--subme 7-10	--rd 4-6	via -preset
Variance-AQ	--aq-mode 1, --aq-strength 1.0	--aq-mode 1-4	--aq-mode 2
Mbtree	--mbtree (on by default)	--cu-tree (on)	auto
Psy-RD	--psy-rd 1.0:0.0	--psy-rd 2.0	--enable-tf 1 (PSY fork)

4. The four questions for any 'AI encoder' pitch

1. Which of the four levers does the AI actually move - partition, mode, scene-cut, or perceptual QP?
2. Is the output bitstream standards-compliant (H.264/HEVC/AV1/VVC), or does it need a custom decoder?
3. What model size and architecture - thousands of parameters or hundreds of millions?
4. What is the BD-rate trade in their own published numbers, and against which baseline?

5. Encoder ASICs that ship these algorithms

Chip	Codecs	Density (1080p60)	Notes
Google Argos VCU	H.264, VP9 (gen 1)	Up to 33x CPU	AV1 + ML on gen 2
NETINT Quadra	H.264, HEVC, AV1	320+ per 1RU	Trillion+ min encoded
NVIDIA NVENC	H.264, HEVC, AV1	Per-GPU varies	Software AQ on top
Apple VideoToolbox	H.264, HEVC, ProRes	Per-SoC	Apple Silicon only

6. Rules of thumb

- Never benchmark a perceptual encoder with PSNR - use VMAF, SSIMULACRA2, or MOS.
- A claimed 30% bitrate save without naming the baseline preset is marketing, not data.
- Production ML models are tiny (1-2 hidden layers, 16-64 nodes), not transformers.
- Scene-cut + partition gate together explain most of the speed gap between encoder presets.
- For 4K/60 fps real-time, a VPU is now cheaper per stream than a CPU farm by ~20x.