RTX 4090 · Optimization Summary

EXP-ISODICE-001
SUMMARY · 2026-06-11

iso-Dice · kidney-tumor CT segmentation (KiTS) · pre-trained nnU-Net (frozen) · successive improvements over the NVIDIA L4 cloud baseline · 10 held-out cases (480–489)

1.0 OPTIMIZATION LADDER vs L4 BASELINE

EACH STEP MEASURED · ISO-DICE EXCEPT WHERE NOTED

CONFIGURATION	TUMOR DICE	s/CASE	AVG W	g CO_2	SPEED vs L4	ENERGY vs L4
BASELINE — CLOUD NVIDIA L4						
L4 baseline (reference) fp32 · TTA · 0.50 — the deployment baseline.	0.887	499	68	3.7	1.0× (ref)	1.0× (ref)
DESKTOP RTX 4090 — SUCCESSIVE OPTIMIZATIONS						
+ half precision (O1) fp16. Lossless — identical Dice.	0.887	123	237	3.2	4.0× faster	1.16× less
+ reduced overlap, no TTA (O3) fp16 · step 0.75 · no TTA. Max efficiency; Dice ↓.	0.780	42	74	0.34	11.8× faster	10.9× less
+ native TensorRT (O4) O3 + compiled fp16 engine. Same Dice.	0.780	≈42	63	0.29	11.8× faster	12.9× less
+ 4-pass TTA (balanced) O3+TRT + light TTA. Recovers ≥0.80.	0.810	46	144	0.73	10.8× faster	5.1× less

TBL-1. Each RTX 4090 row is measured against the cloud L4 baseline (top, reference) over the same 10 held-out cases. Avg W = mean GPU power (energy ÷ wall, NVML/Zeus). g CO_2 = GPU energy × 393.5 g/kWh (the US-grid intensity CodeCarbon measured this run), applied uniformly for comparability — it scales with the deployment grid (≈0 on hydro/geothermal, higher on coal); CodeCarbon's full GPU+CPU+RAM figure is roughly 2× these. The balanced row's energy (5.1× less than L4), Dice and CO_2 are exact — energy is anchored to O3's measured value via an in-harness energy ratio; its wall, power and speed are anchored to O3's official figures (±~10 %). Half precision (O1) is lossless; O3 is maximum efficiency at a 0.107 Dice cost; the native TensorRT engine (O4) holds that Dice with less energy; 4-pass TTA (balanced) restores tumor Dice to 0.810.

CHOOSE BY PRIORITY

Full accuracy	O1 — fp16, tumor Dice 0.887 (= baseline), 4× faster / 16 % less energy than the L4.
Balanced (≥0.80)	O3 + TensorRT + 4-pass TTA — tumor Dice 0.810, 5.1× less energy than the L4 (0.73 g CO_2/case at US grid).
Max efficiency	O3 + TensorRT — tumor Dice 0.780, 12.9× less energy than the L4 (use only if 0.78 clears the clinical threshold).
Runtime add-on	all_in_gpu = True — a lossless ~1.74× throughput / 15 % energy win on any of the above (single flag, +3 GB VRAM).

2.0 GLOSSARY

PLAIN-LANGUAGE TERMS

Tumor Dice — overlap between predicted and reference tumor segmentation (0–1; higher is better). The limiting accuracy metric.

fp16 / fp32 — half- vs single-precision arithmetic. Lower precision runs faster and uses less energy, and here is lossless.

iso-Dice / lossless — the optimization produces the same segmentation; only speed, energy or memory change, not accuracy.

TTA (test-time augmentation) — averaging the prediction over mirrored copies; more accurate, at 2× the compute per mirrored axis (1 → 2 → 4 → 8 passes).

Patch step / overlap — the sliding window scans in patches; step 0.50 = 50 % overlap, step 0.75 = 25 % overlap (fewer patches, faster, coarser).

all_in_gpu — keeps the patch-by-patch aggregation on the GPU instead of copying each patch to the CPU; a lossless throughput setting.

CO_2 (CodeCarbon) — energy × the local grid's carbon intensity (g CO_2 per kWh); CodeCarbon also adds modelled CPU + RAM. Depends heavily on where the GPU runs.

TensorRT — NVIDIA's inference compiler; builds an optimized fp16 engine ("native" = built from an ONNX export).

NVML / Zeus — the GPU's on-chip hardware energy counter (Zeus wraps the same counter); the authoritative kJ/case and W figures.

L4 / RTX 4090 — a cost-efficient cloud datacenter GPU (the deployment baseline) vs a desktop/consumer GPU.

Energy/power from NVML + Zeus hardware counters (CodeCarbon/CarbonTracker soft meters are disabled on these pods — they crash on blocked RAPL; CodeCarbon was run separately for CO_2 only). Abandoned experiments (torch-tensorrt fp16 collapse, connected-component postprocessing, multi-process concurrency, INT8) are omitted; see the full report.