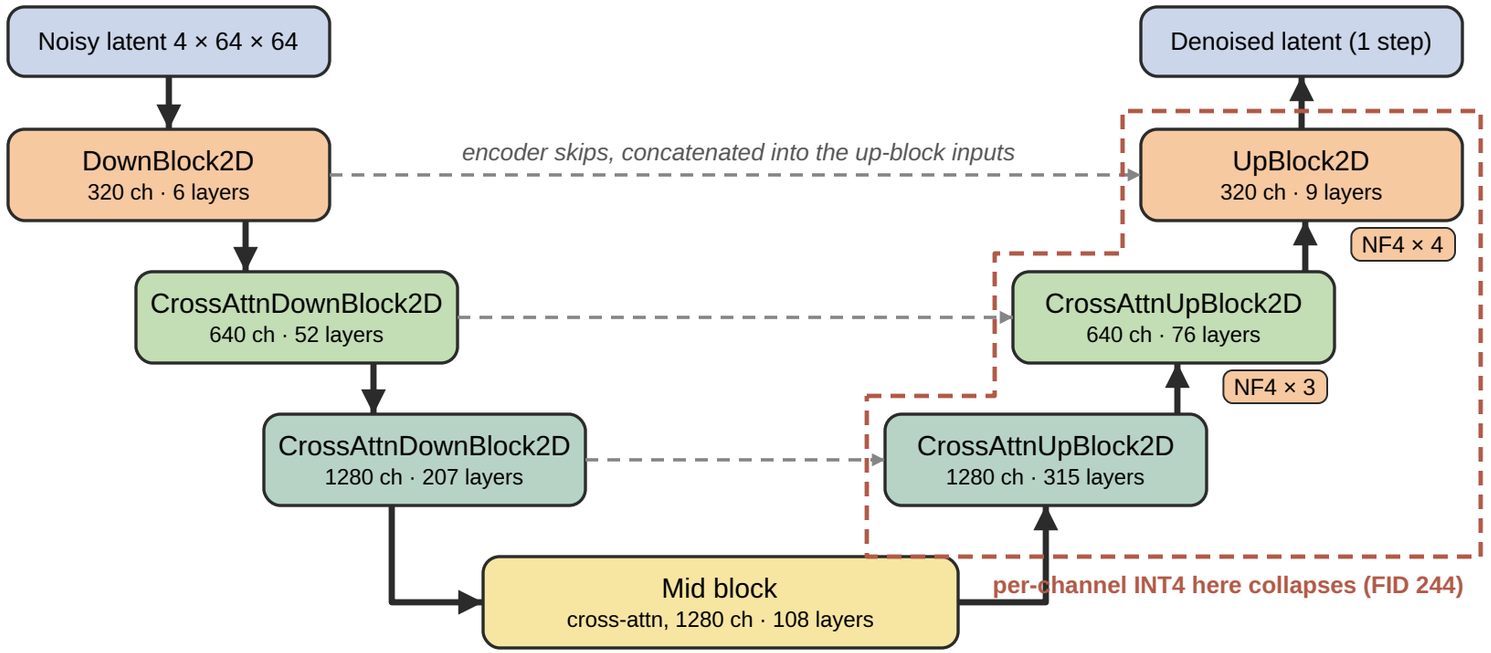
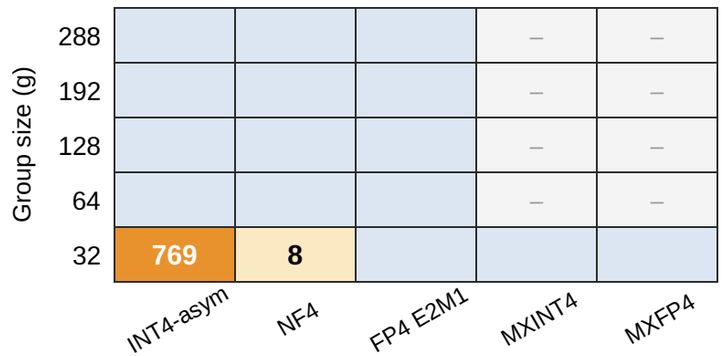
SDXL-Turbo: one denoising step at 512 × 512, no CFG



W_4 Weight Quantization (Offline)

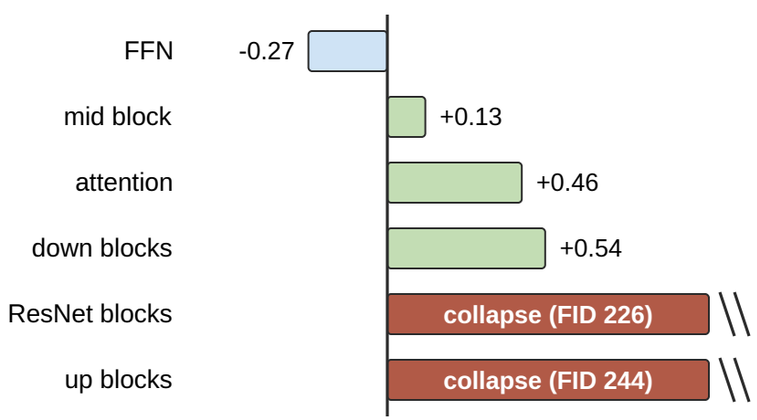
Per-layer joint (g, f) search, every layer at 4 bits.



769 of 777 layers (99%): INT4-asym at g = 32;
8: NF4 at g = 32 (7 in the up-blocks); FP4 and MX never chosen.
Block counts exclude the 4 time/added-embedding linears.
INT4-sym is not in the few-step palette; MX formats use 32-blocks (–).
The 9 up-block conv_shortcut layers stay FP16.

Layer Sensitivity

one block group → per-output-channel INT4 (still 4-bit)

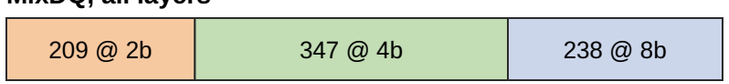


ΔFID vs Chameleon (26.72, 5k)

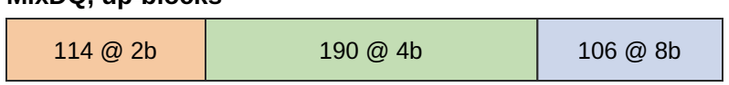
Bits per Layer at W_4

MixDQ mixes bit-widths; Chameleon keeps 4 bits

MixDQ, all layers



MixDQ, up-blocks



Chameleon, all layers



2-bit 4-bit 8-bit

Table 1, W_4A_8 FID: MixDQ 24.50, Chameleon 21.29

A_8 Activations (Runtime)

- one step: no timestep axis, no format LUT
- MixDQ's calibrated static scales and mixed-precision assignment, unchanged
- MixDQ's BOS-aware bypass dropped: batched inference restored
- Chameleon owns the weights only