Timestep t

Size / crop ids

Time + added embedding
(t, pooled text, size/crop)

pooled text

CLIP text embedding
(2048-d)

text encoders (FP16)

K/V

timestep conditioning

Noisy VAE latent,
4 × 128 × 128

conv_in

FP16

DownBlock2D
(320 ch)

CrossAttnDownBlock2D
(640 ch)

CrossAttnDownBlock2D
(1280 ch)

Mid block
cross-attn, 1280 ch

c

conv1 ×3

c

conv1 ×3

c

conv1 ×3

CrossAttnUpBlock2D
(1280 ch)

CrossAttnUpBlock2D
(640 ch)

UpBlock2D
(320 ch)

conv_out

FP16

Predicted noise
(epsilon-hat)

A_8 Activation Quantization (Runtime)

Macro-routing, static scales

share of the 372 layers per format, per bucket

Timestep Bucket → Format Code

clean signal (low t) → pure noise (high t)

Bucket	INT8	E4M3	E5M2
T1	10%	**50%**	37%
T2	12%	**48%**	37%
T3	12%	**49%**	37%
T4	–	**61%**	36%
T5	–	**62%**	35%
T6	–	**62%**	36%
T7	–	–	**98%**
T8	–	–	**98%**
T9	–	–	**98%**
T10	–	–	**98%**

MXFP8: the 9 shortcut `conv1` (2.4%) in every bucket
per-bucket scales, calibrated ahead of time
and static at inference (Q-Diffusion style)

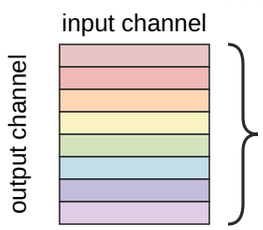
conv1 Shortcut Topology Rule

`up_blocks.*.resnets.*.conv1`
→ MXFP8 E4M3, unconditionally, bypassing the κ/SNR check; ×3 per level, one per up-block ResNet.

The concatenated input spans a peak-to-median channel range of 58× at the median layer and up to 163× (~1.4× an ordinary conv).

W_4/W_8 Weight Quantization (Offline)

SDXL: per-output-channel format selection.



W8: {INT8-sym, INT8-asym, MXINT8}

W4: {INT4-sym, INT4-asym, NF4,
FP4 E2M1, MXINT4, MXFP4 E2M1}

372 quantized conv and linear layers; attention Q/K/V and `conv_in`/`conv_out` stay FP16.

Few-Step Variant (SDXL-Turbo, Figure 4)

- same UNet topology; 1 denoising step at 512 × 512
- weights: per-layer (g, f) search, 99% INT4-asym g = 32
- activations: MixDQ static scales + mixed precision
- MixDQ's BOS-aware bypass dropped
- concat: 9 up-block `conv_shortcut` layers stay FP16; no MXFP8 pin

c Skip Concatenation

Channel-wise `torch.cat` of the decoder stream (from below) with one encoder skip, before each of the three up-block ResNets.

Per level, the skips are the two ResNet outputs of the same-level encoder block and the downsampled output of the level above (`conv_in`'s output at the top).

Both `conv1` (pinned to MXFP8) and the parallel 1×1 `conv_shortcut` (routed by κ/SNR) read its output.