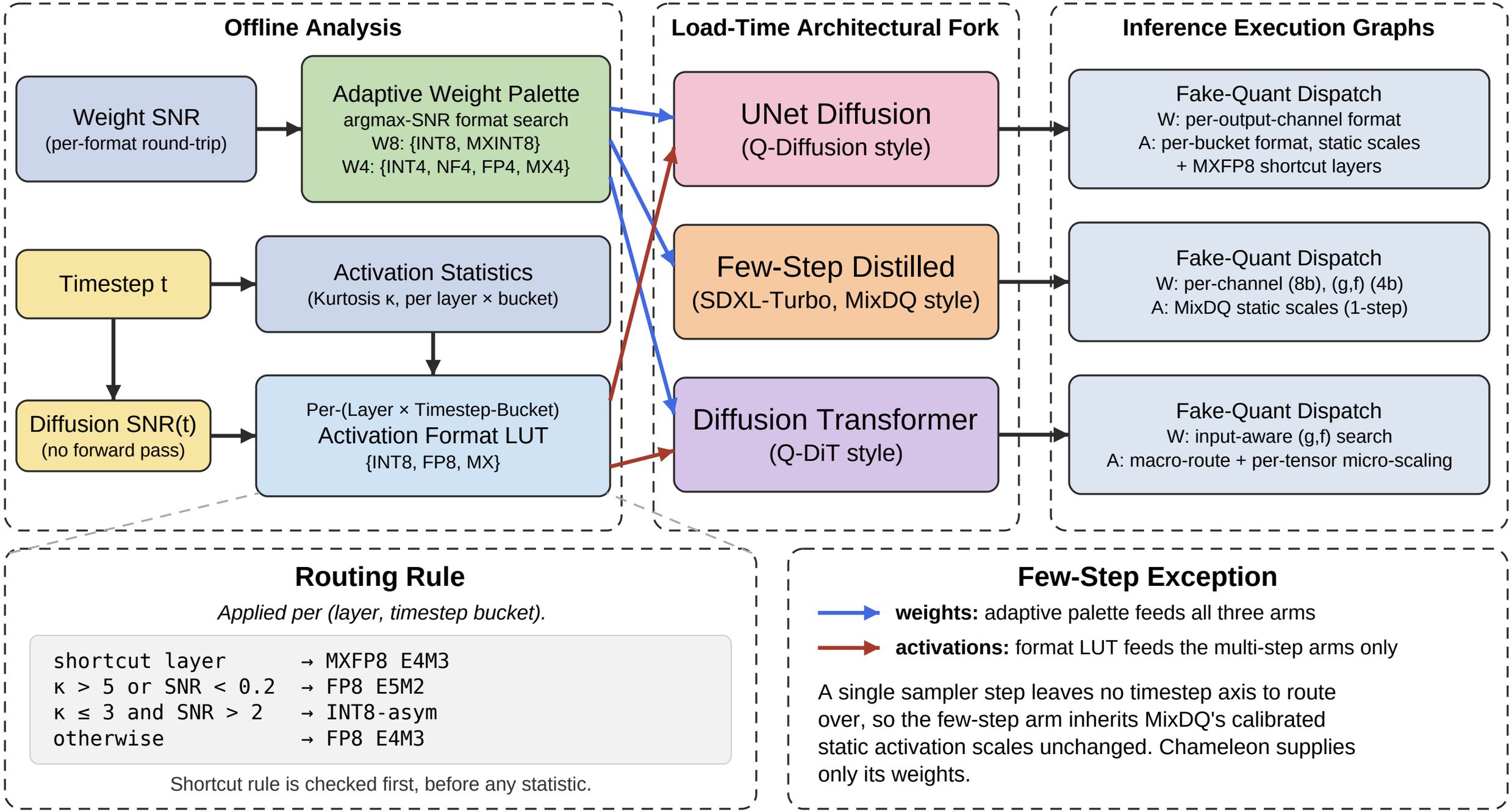

Offline Analysis
Weight SNR
(per-format round-trip)
Adaptive Weight Palette
argmax-SNR format search
W8: {INT8, MXINT8}
W4: {INT4, NF4, FP4, MX4}
Timestep t
Activation Statistics
(Kurtosis κ, per layer × bucket)
Diffusion SNR(t)
(no forward pass)
Per-(Layer × Timestep-Bucket)
Activation Format LUT
{INT8, FP8, MX}
Load-Time Architectural Fork
UNet Diffusion
(Q-Diffusion style)
Few-Step Distilled
(SDXL-Turbo, MixDQ style)
Diffusion Transformer
(Q-DiT style)
Inference Execution Graphs
Fake-Quant Dispatch
W: per-output-channel format
A: per-bucket format, static scales
+ MXFP8 shortcut layers
Fake-Quant Dispatch
W: per-channel (8b), (g,f) (4b)
A: MixDQ static scales (1-step)
Fake-Quant Dispatch
W: input-aware (g,f) search
A: macro-route + per-tensor micro-scaling
Routing Rule
Applied per (layer, timestep bucket).
shortcut layer → MXFP8 E4M3
κ > 5 or SNR < 0.2 → FP8 E5M2
κ ≤ 3 and SNR > 2 → INT8-asym
otherwise → FP8 E4M3
Shortcut rule is checked first, before any statistic.
Few-Step Exception
weights: adaptive palette feeds all three arms
activations: format LUT feeds the multi-step arms only
A single sampler step leaves no timestep axis to route over, so the few-step arm inherits MixDQ's calibrated static activation scales unchanged. Chameleon supplies only its weights.