Noisy VAE latent,
4 × 128 × 128

Patchify
(patch size 2)

Sequence:
4096 tokens
× 1152 dim

Timestep t

Timestep embedding

AdaLN-single
(t, resolution, aspect ratio)

`adaln_single` *(FP16)*

one global MLP, shared by
all 28 blocks; summed with
each block's `scale_shift_table`

DiT Block **× 28**

LayerNorm + modulate

Self-Attention
(attn1)

× `gate_msa`

K/V

Cross-Attention
(attn2)

LayerNorm + modulate

Feed-Forward
(MLP)

× `gate_mlp`

Caption projection

T5 text embedding
(4096-d, 120 tokens)

Final LayerNorm
+ modulate

Linear
(proj_out)

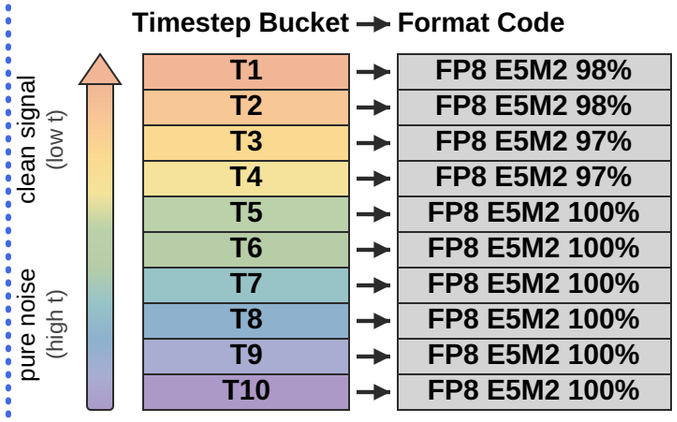
Unpatchify

Predicted
noise
(epsilon-hat)

A_8 Activation Quantization (Runtime)

Macro-routing

format code per (layer, timestep bucket)



one LUT read per layer per step

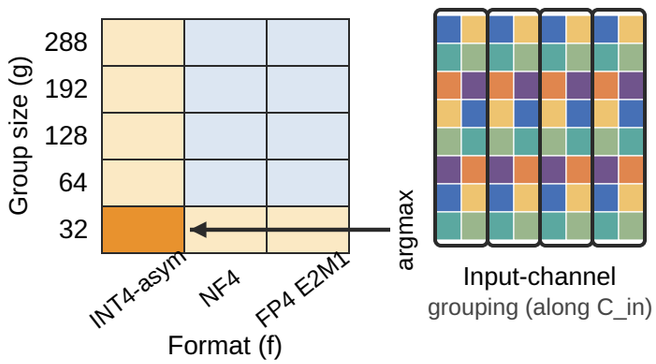
INT8-asym never selected: DiT activations stay heavy-tailed at every timestep, so the table is effectively static

Micro-scaling

scale computed dynamically
from the current tensor's min/max at inference

W_4/W_8 Weight Quantization (Offline)

Per-linear joint (g, f) search (input-aware at W_4).



W_4: g = 32 for all 282 layers;
INT4-asym 213, NF4 69, FP4 E2M1 0
W_8: INT8-asym at g = 32 everywhere (MXINT8 never)

Real kernel

one Triton kernel decodes all three formats
through a 32-entry LUT (Ada, exact numerics)

282 quantized linears (28 × 10 + caption projection);
`proj_out` and `adaln_single` stay FP16.