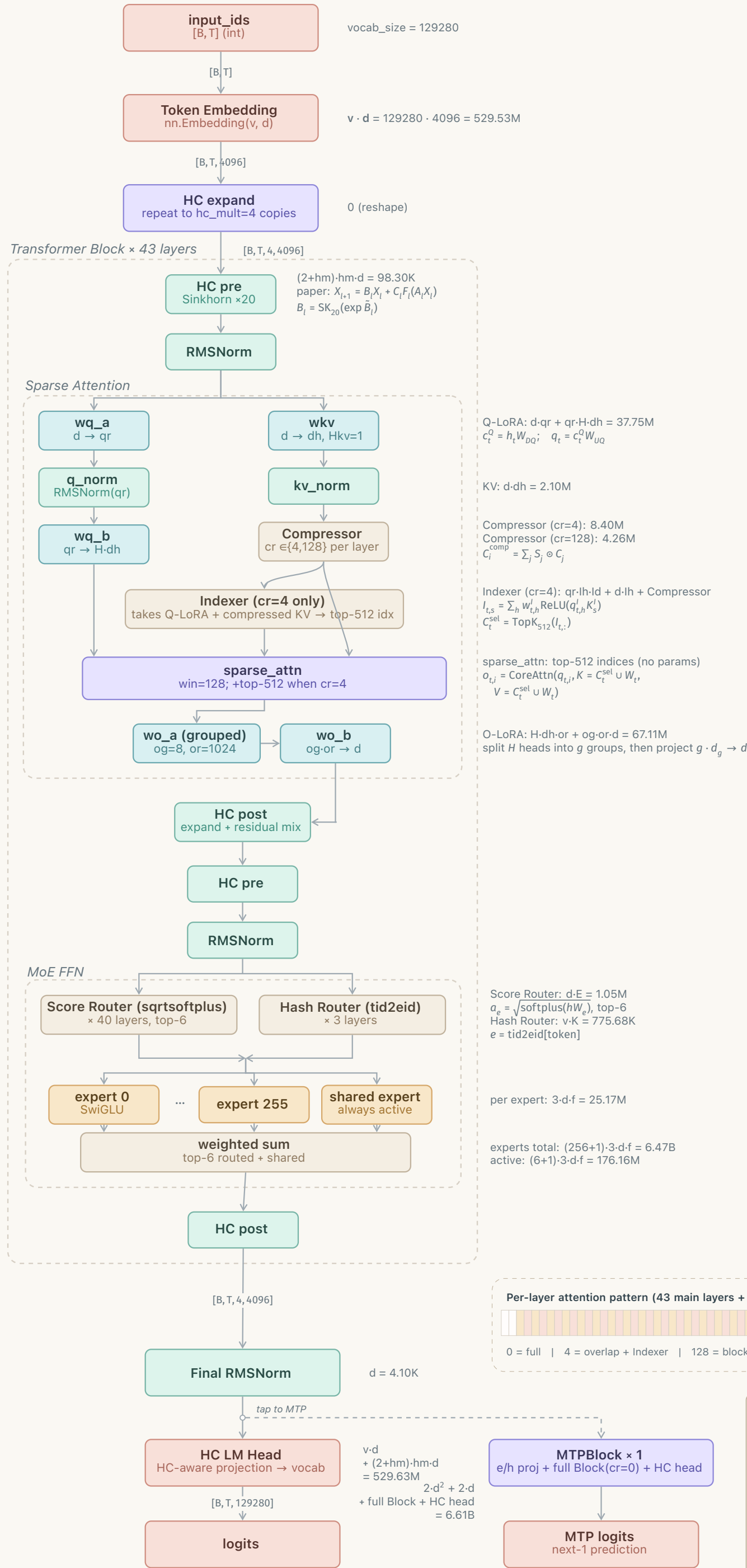


deepseek-ai/DeepSeek-V4-Flash

DSV4 family · 284.44B total · 13.90B active per token · hash_routing, moe, mtp, shared_experts, sliding_window, swiglu_clip



Model	deepseek-ai/DeepSeek-V4-Flash
Family	dsv4 (deepseek_v4)
Total params	284.44B (main stack)
Active per token	13.90B
MTP head	6.61B
Total + MTP	291.05B
Quantization	FP8, e4m3, experts=fp4, BF16 gates/norms
Layers (L)	43
Hidden (d)	4096
Heads (H/Hkv)	64 / 1
Head dim (dh)	512
Q LoRA rank	1024
O LoRA / groups	1024 / 8
Routed experts	256 (top-6)
Shared experts	1
MoE FFN dim (f)	2048
HC mult (hm)	4
Index top-K	512
Sliding window	128