


模型参数定义

ZOMI


Contenet


Transformer 架构
Transformer 前世今生 3
核心机制之 Token 3
核心机制之 Embedding 5
核心机制之 Attention
核心机制之 Attention变种 2
长序列挑战 4
大模型参数配置 & 比例


好好想想

视频目录大纲

- Q/K/V维度与头数配置
- 不同的模型参数
- 模型结构差异

01

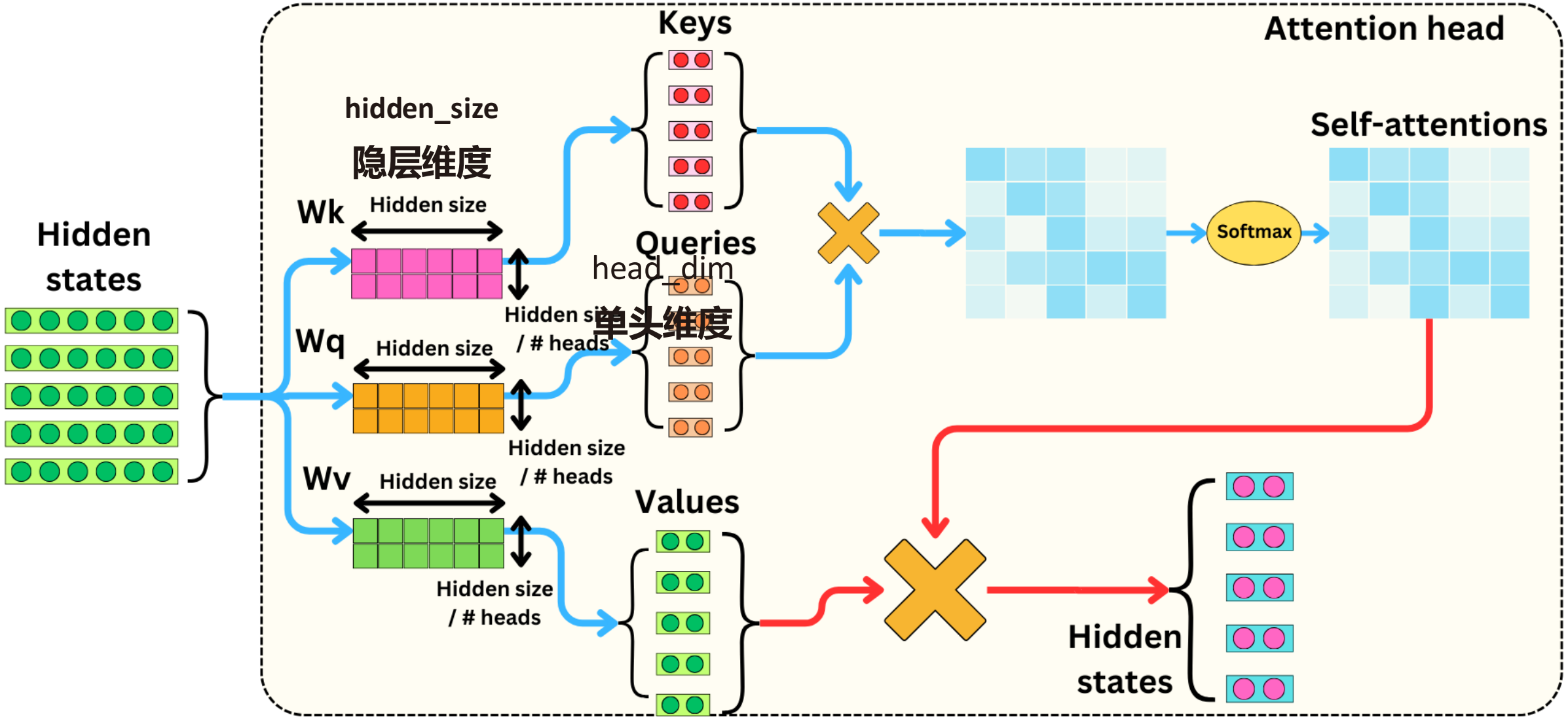
Q/K/V维度与头数配置

注意力头数（Head Num）

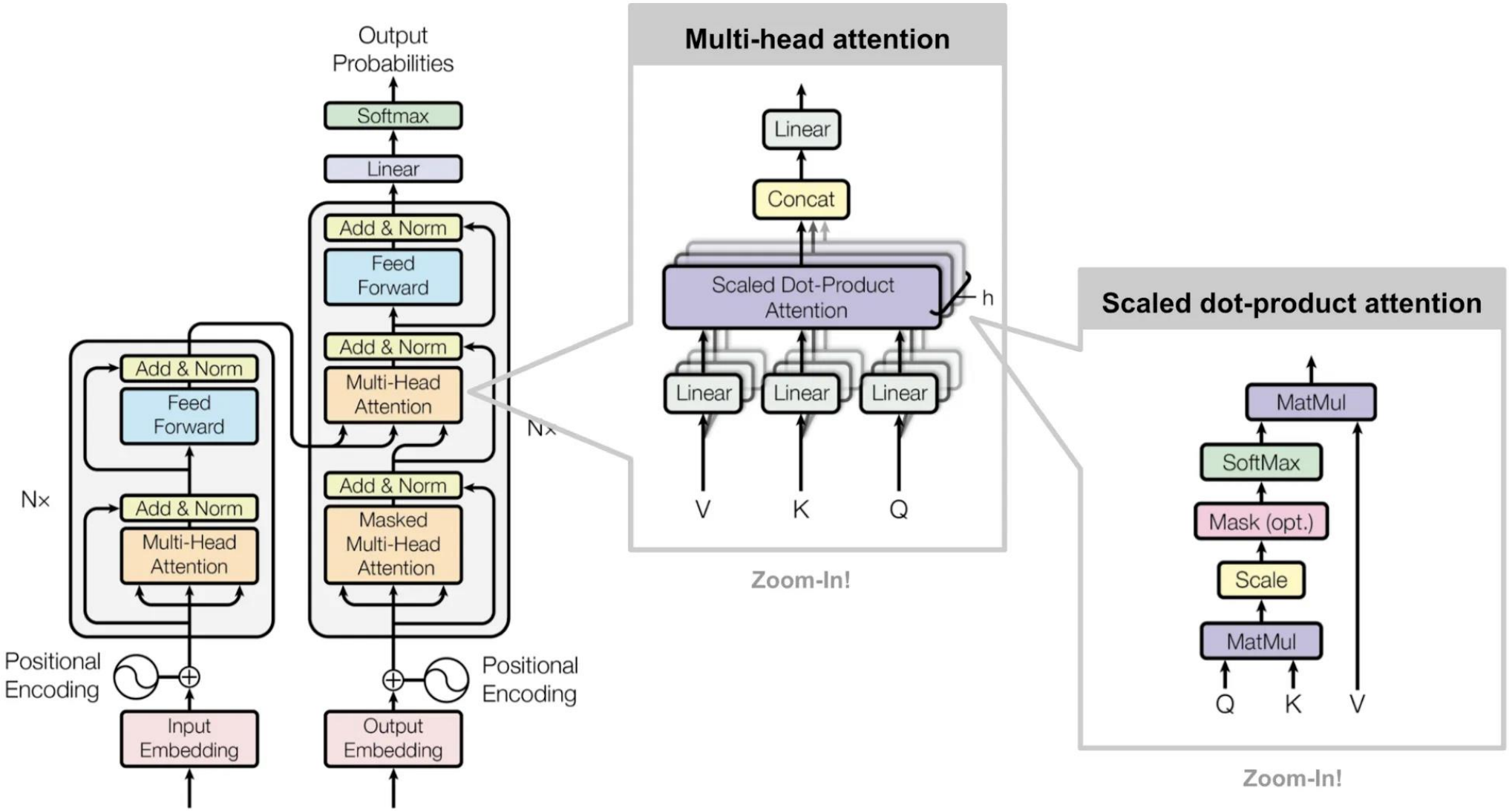
- 与隐层维度（hidden_size）成比例，单头维度为 hidden_size / num_heads
  - Qwen2-7B：隐层维度 3584，头数 28，单头维度为 3584/28=128
  - Llama3-8B：隐层维度 4096，头数 32，单头维度为 4096/32=128

- **原则：**保持单头维度在 64-128 之间，确保 Transformer 架构特征捕捉能力与计算效率平衡

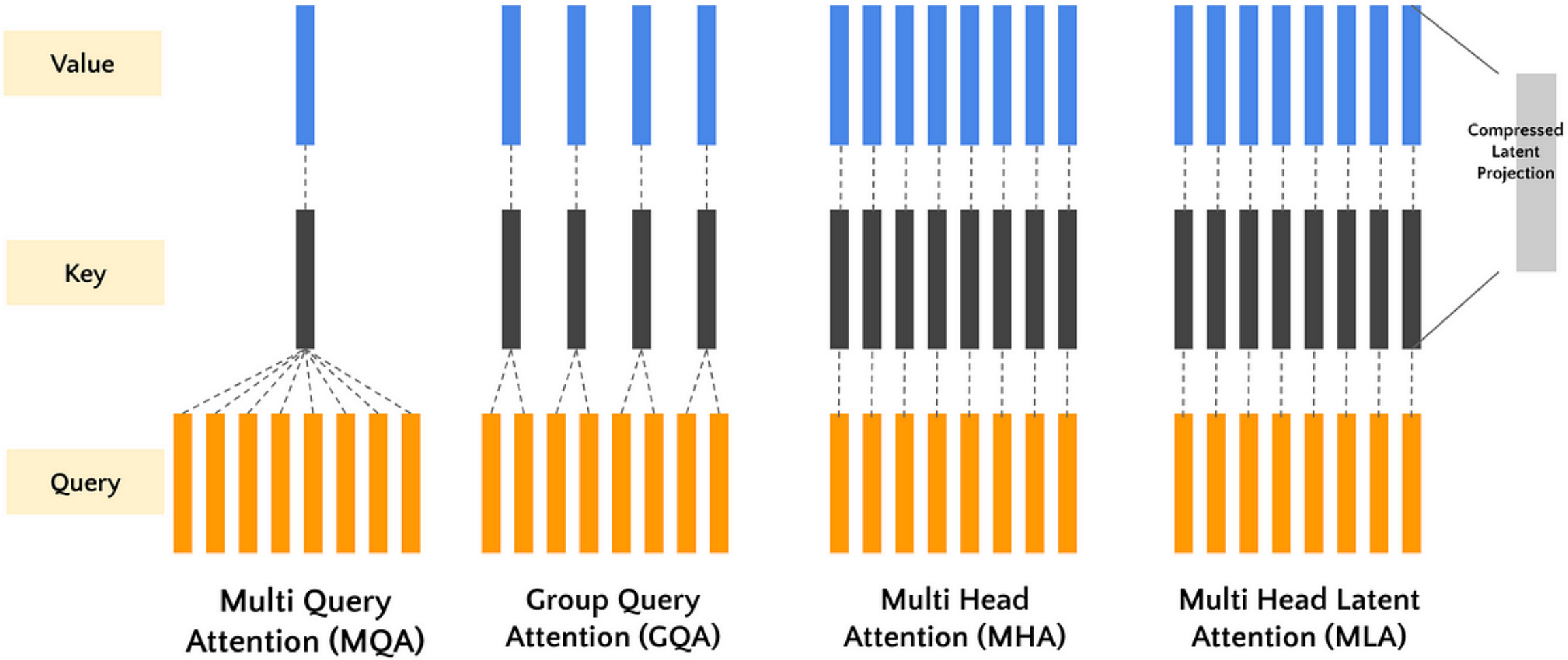
注意力头数（Head Num）



注意力头数 num_heads



关键值头数（Key-Value Heads）



关键值头数（Key-Value Heads）

- 分组查询注意力（GQA）中，KV 头数可减少以降低显存消耗：
  - Qwen2-7B：Query 28，Key Value Head 4，形成 7:1 分组比例
  - Llama3-8B：Query 32， Key Value Head 8，形成 4:1 分组比例

- **原则：** Key Value Head 通常为 Query 1/4~1/8，显存需求可减少 30%-50%

参数配置比例

- **头数与维度的平衡：**
  - 增加 head 可提升模型并行计算能力，但需降低 head_dim 以控制显存占用（64~128）。
  - 若 head_dim 过小（<64），可能导致模型表达能力不足。
- **层数扩展：**
  - 堆叠层数每增加一倍（32 to 64），模型参数量和计算量显著上升。
  - 需配合更大的训练数据量（~1.4T tokens）。

02

模型参数

Qwen3-0.6B/1.7B/4B/8B/14B/32B

Model	head dim	hidden act	hidden size	intermediate size	max position embeddings	max window layers	attention heads	num hidden layers	num kv heads	vocab size
Qwen3-0.6B	128	silu	1024	3072	40960	28	16	28	8	151936
Qwen3-1.7B	128	silu	2048	6144	40960	28	16	28	8	151936
Qwen3-4B	128	silu	2560	9728	40960	36	32	36	8	151936
Qwen3-8B	128	silu	4096	12288	40960	36	32	36	8	151936
Qwen3-14B	128	silu	5120	17408	40960	40	40	40	8	151936
Qwen3-32B	128	silu	5120	25600	40960	64	64	64	8	151936

Qwen3-30B-A3B/Qwen3-235B-A22B

Model	head dim	hidden act	hidden size	interme diate size	max position embedd ings	max window layers	moe interme diate size	attentio n heads	num experts	num experts per tok	n_share d_exper ts	num hidden layers	num kv heads	vocab size
Qwen3-30B-A3B	128	silu	2048	6144	40960	48	768	32	128	8	/	48	4	151936
Qwen3-235B-A22B	128	silu	4096	12288	40960	94	1536	64	128	8	/	94	4	151936
DeepSeek-V2-236B	128	silu	5120	12288	163840	/	1536	128	160	6	2	60	128	102400
DeepSeek-V3-671B	128	silu	7168	18432	163840	/	2048	128	256	8	1	61	128	129280

多头注意力（MHA）相关参数

- 头数（num_attention_heads）：
  - 决定模型并行关注不同子空间的能力，头数越多可提取的特征多样性越高，但会增加计算量。
- Q/K/V 线性层参数：
  - Query、Key、Value 向量的投影层参数，直接影响注意力计算的稳定性与效率，需与隐藏层维度（hidden_size）匹配。

FNN 中间层维度

- FFN层的中间维度 intermediate_size 通常为隐藏层维度 hidden_size 的数倍（如4倍），控制模型非线性表达能力。
- 例如LLaMA-2中该值为11008（hidden_size=4096），过大会增加计算成本，过小可能限制模型容量

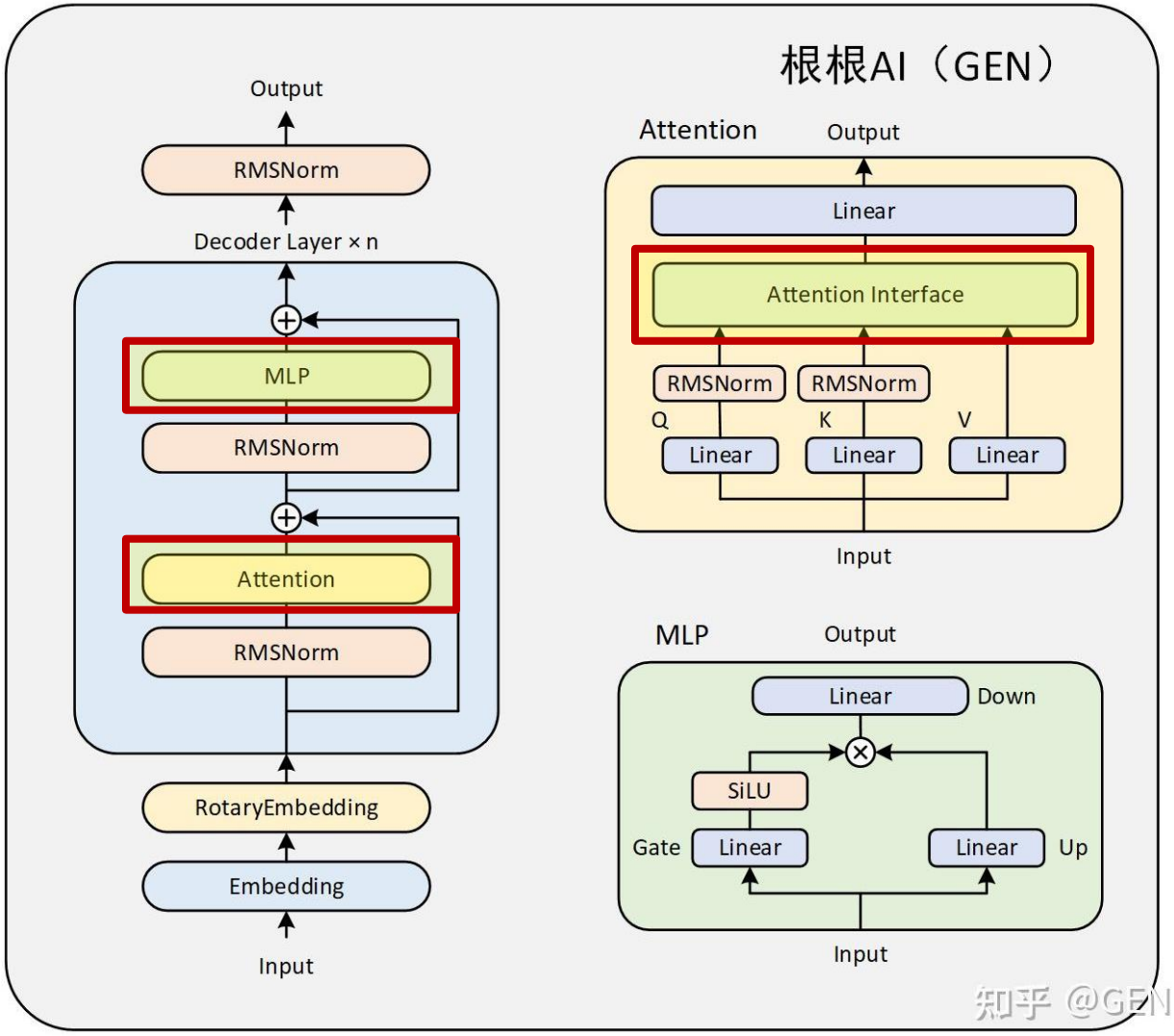
模型深度

- 模型层数决定堆叠的 Transformer 模块数量，深层模型可提升表达能力，但可能导致训练困难或过拟合。需结合任务复杂度与硬件资源权衡。

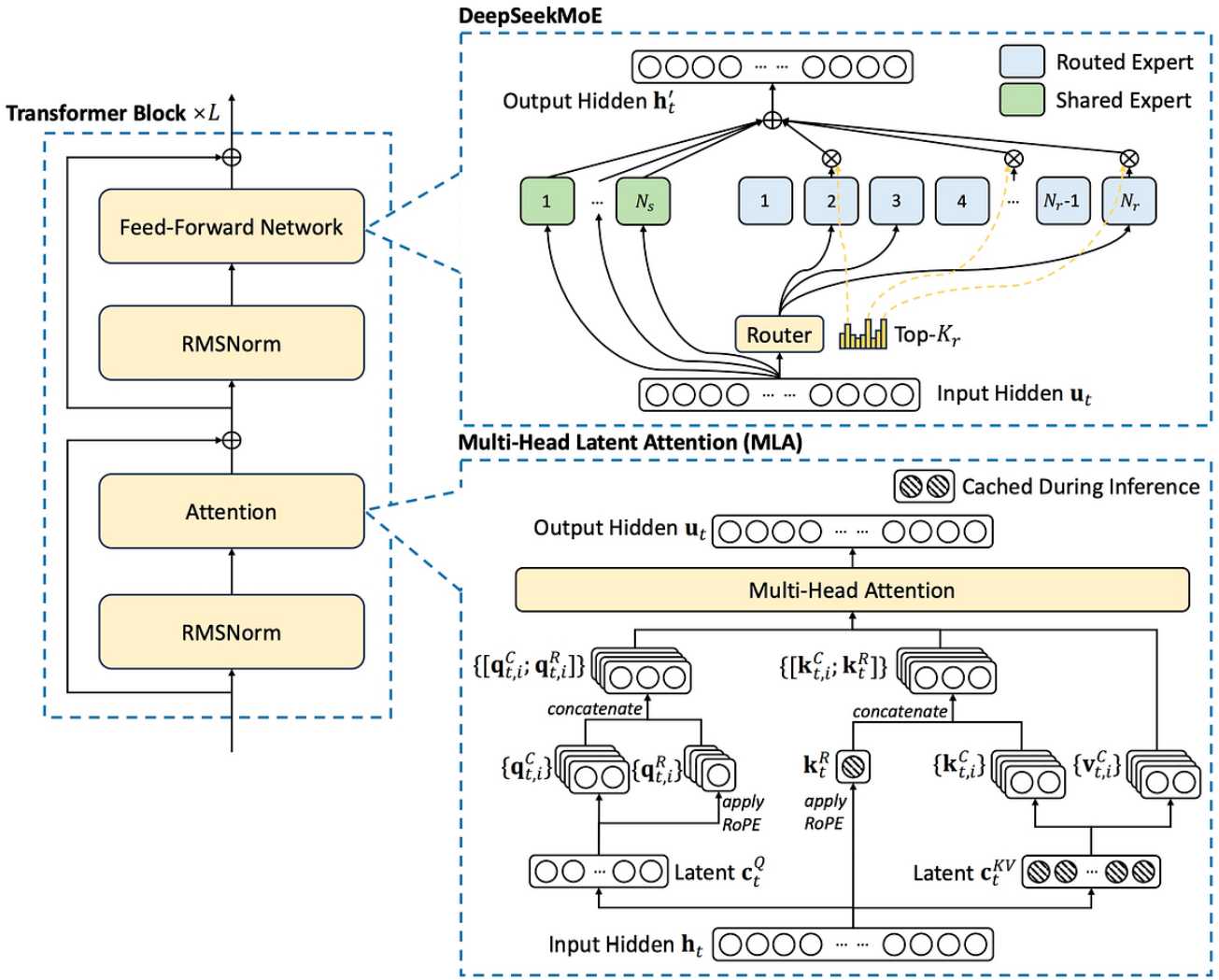
03

模型结构差异

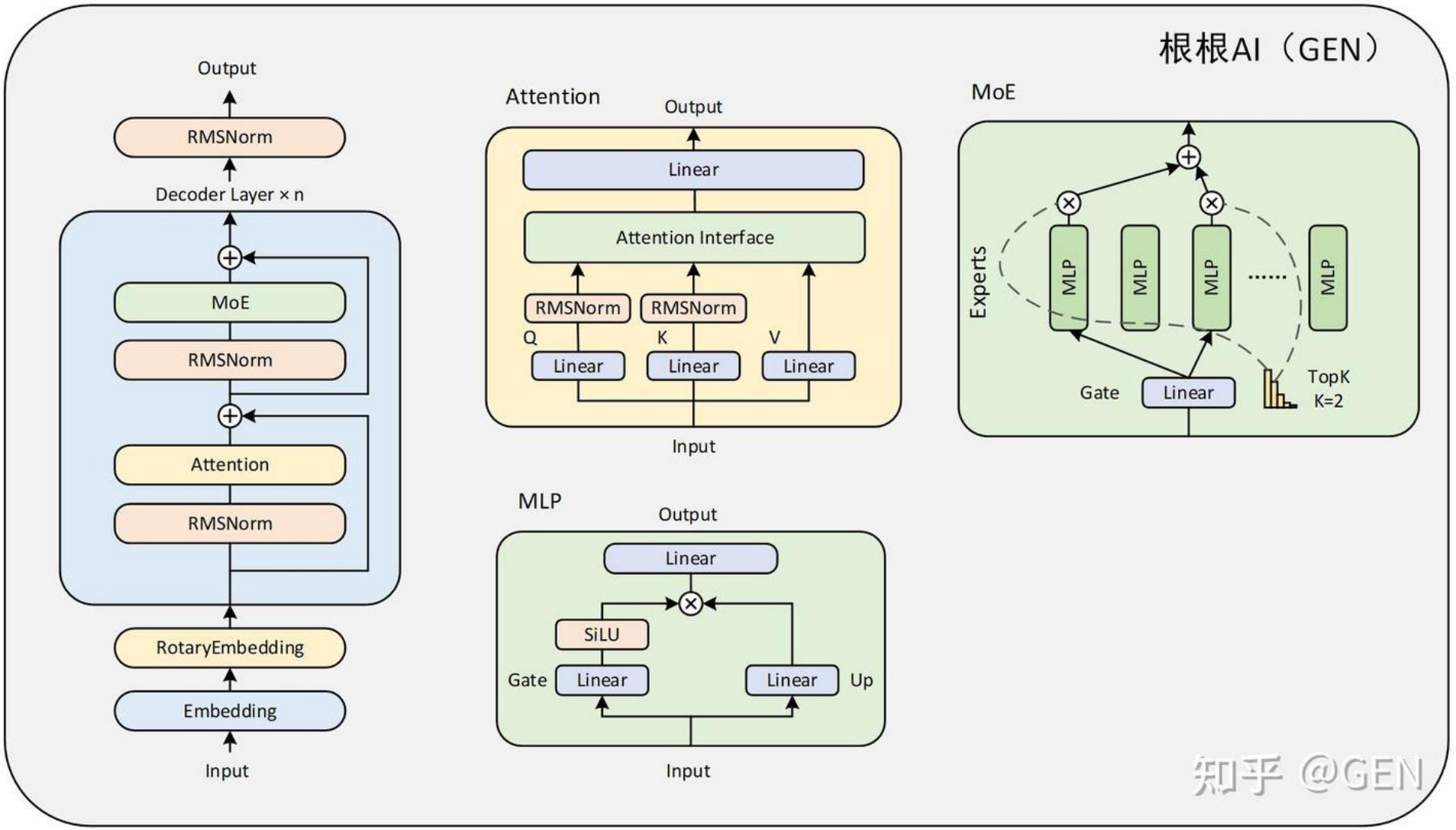
Qwen3 Dense


根根AI（GEN）
Output
RMSNorm
Decoder Layer × n
MLP
RMSNorm
Attention
RMSNorm
RotaryEmbedding
Embedding
Input
Attention
Output
Linear
Attention Interface
RMSNorm
RMSNorm
Q
K
V
Linear
Linear
Linear
Input
MLP
Output
Linear
Down
SiLU
Gate
Linear
Linear
Up
Input
知乎 @GEN

DeepSeek


Transformer Block ×L
Feed-Forward Network
RMSNorm
Attention
RMSNorm
DeepSeekMoE
Output Hidden $\mathbf{h}'_t$
Routed Expert
Shared Expert
1
N_s
1
2
3
4
N_r-1
N_r
Router
Top-K_r
Input Hidden $\mathbf{u}_t$
Multi-Head Latent Attention (MLA)
Cached During Inference
Output Hidden $\mathbf{u}_t$
Multi-Head Attention
$\{[\mathbf{q}^C_{t,i}; \mathbf{q}^R_{t,i}]\}$
$\{[\mathbf{k}^C_{t,i}; \mathbf{k}^R_t]\}$
concatenate
concatenate
$\{\mathbf{q}^C_{t,i}\}$
$\{\mathbf{q}^R_{t,i}\}$
apply RoPE
$\mathbf{k}^R_t$
apply RoPE
$\{\mathbf{k}^C_{t,i}\}$
$\{\mathbf{v}^C_{t,i}\}$
Latent $\mathbf{c}^Q_t$
Latent $\mathbf{c}^{KV}_t$
Input Hidden $\mathbf{h}_t$

Qwen3 MoE


根根AI（GEN）
Output
RMSNorm
Decoder Layer × n
MoE
RMSNorm
Attention
RMSNorm
RotaryEmbedding
Embedding
Input
Attention
Output
Linear
Attention Interface
RMSNorm
RMSNorm
Q
K
V
Linear
Linear
Linear
Input
MoE
Experts
MLP
MLP
MLP
.....
MLP
Gate
Linear
TopK
K=2
MLP
Output
Linear
SiLU
Gate
Linear
Linear
Up
Input
知乎 @GEN

Qwen3 vs DeepSeekV3

维 度	Qwen3 MoE	DeepSeek-V3
专家结构	128 个专家，每个 Token 激活 8 个专家，细粒度分割提升表达效率	671B 总参数，每个 Token 激活 37B 参数，专家负载动态调整
注意力机制	引入 QK-Norm 优化注意力稳定性，支持滑动窗口上下文（4096 Token窗口）	采用多头潜在注意力 MLA，通过低秩压缩减少 70% KV 缓存，支持超长序列推理
路由策略	全局批次负载均衡损失（Global-Batch Load Balancing Loss），无共享专家设计	动态路由偏置项，无需辅助损失函数实现专家负载均衡，降低训练复杂度
位置编码	动态调整 RoPE 频率，支持 YARN 和 Dual Chunk Attention 扩展至 128K 上下文	固定 RoPE 扩展策略，结合 MLA 优化长序列处理效率

总结与思考

总结与建议

- 优先调优顺序：层数 → 头数 → 隐藏层 → 初始化 → 正则化
- 避坑指南：避免层数/头数盲目堆砌，需通过验证集性能监控动态调整。

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

GitHub https://github.com/chenzomi12/AIInfra

引用与参考

- https://www.youtube.com/watch?v=sOPDGQjFcuM&t=1s
- https://arxiv.org/abs/2308.00951
- https://arxiv.org/abs/2106.05974
- https://zhuanlan.zhihu.com/p/652536107

- PPT 开源在：https://github.com/chenzomi12/AIInfra