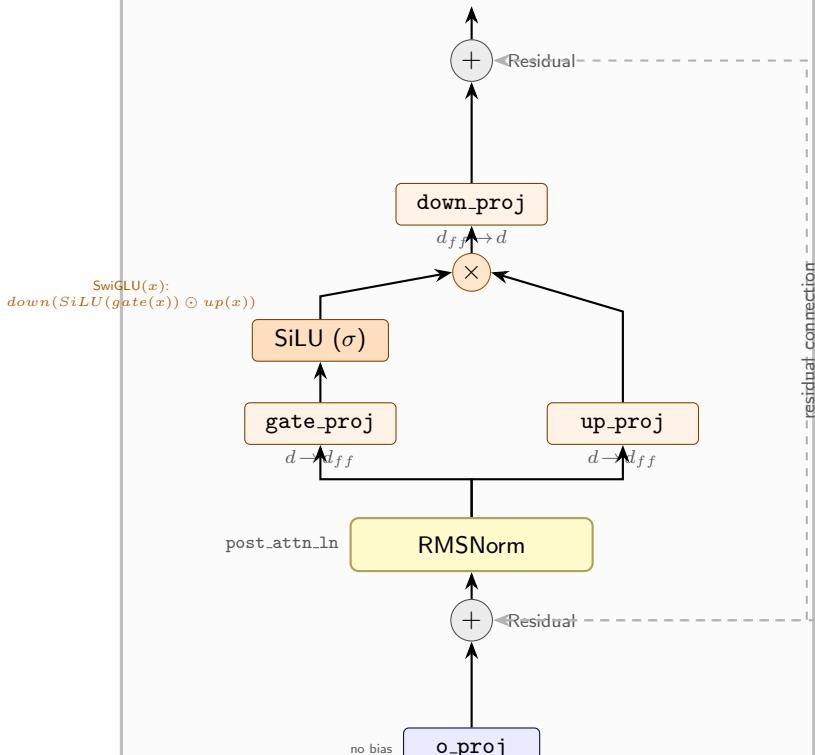
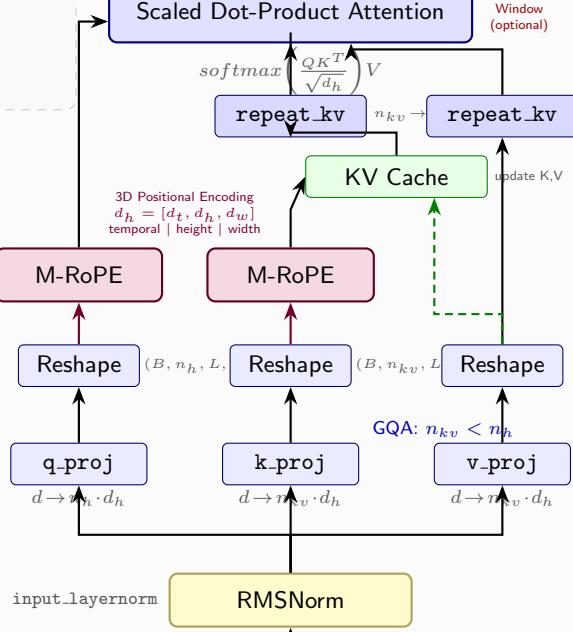


Output Hidden States (B, L, d)



LLM Decoder vs Vision Encoder:

- Separate Q/K/V proj \rightarrow Combined QKV
- GQA ($n_{kv} < n_h$) \rightarrow MHA ($n_{kv} = n_h$)
- M-RoPE (3D pos) \rightarrow 2D RoPE
- SwiGLU MLP \rightarrow Simple MLP
- Sliding Window \rightarrow Full Attention
- KV Cache \rightarrow No Cache
- Causal mask \rightarrow No causal mask



Qwen2VLDecoderLayer ($\times N$ layers)