Qwen2-VL VisionBlock Internal Structure

Qwen2VLVisionBlock: Pre-Norm Transformer with 2D RoPE and Window Attention

Input hidden_states $[N, 1152]$

$eps = 10^{-6}$

residual

LayerNorm1 $[N, 1152]$

VisionAttention

$+$ $[N, 1152]$

$eps = 10^{-6}$

residual

LayerNorm2 $[N, 1152]$

VisionMlp

$+$ $[N, 1152]$

Output hidden_states $[N, 1152]$

VisionAttention (Expanded)

input: $[N, 1152]$

$bias = True$ **QKV Linear** $1152 \rightarrow 3456$

reshape + unbind

Q $[N, 16, 72]$ **K** $[N, 16, 72]$ **V** $[N, 16, 72]$

$position\_embeddings$ **2D RoPE** $\rightarrow$ **2D RoPE** *(no RoPE)*

Q' K'

$cu\_seqlens$ $\rightarrow$ **Window Attention**

num_heads=16, head_dim=72

$[N, 1152]$

Output Proj $1152 \rightarrow 1152$

output: $[N, 1152]$

VisionMlp (Expanded)

input: $[N, 1152]$

$\times 4$ *expansion* **fc1 Linear** $1152 \rightarrow 4608$

QuickGELU $[N, 4608]$

fc2 Linear $4608 \rightarrow 1152$

output: $[N, 1152]$

mlp_ratio=4

Legend

LayerNorm   Attention   MLP   RoPE   Residual