

Qwen2.5-VL and RoPE: Oral Knowledge Questions

Explain each answer in your own words and justify the underlying mechanism.

1. **The PatchMerger adapter.** Qwen2.5-VL's adapter combines each group of 2×2 neighboring visual patch tokens into one token for the language decoder.

Explain how the four patch features are normalized, concatenated, and transformed by a two-layer MLP. If each input feature has dimension 1280, what is the concatenated vector's dimension, and what is the output dimension for Qwen2.5-VL-7B? How does this operation change the number of visual tokens passed to the decoder?

2. **Attention in the ViT and language decoder.** Compare window attention in Qwen2.5-VL's vision encoder with causal self-attention in its language decoder. Which tokens can interact in each case? Explain how the full-attention ViT layers provide global image context and how an answer token can use visual information from different image regions.

3. **What RoPE rotates.** Explain how RoPE incorporates position into attention. Which vectors are rotated, how is a higher-dimensional vector divided into two-dimensional pairs, and why are different rotation frequencies used? How does this differ from adding a positional vector to the input token embedding?

4. **Absolute positions and relative attention.** RoPE rotates a query using its absolute position m and a key using its absolute position n . Why does their dot product nevertheless involve the relative offset $n - m$? Explain what the rotation preserves and why the resulting attention score still depends on token content.

5. **M-RoPE for text, images, and video.** How does M-RoPE assign temporal, height, and width coordinates to text tokens, image tokens, and video tokens?

Explain what a **constant temporal coordinate within one image** means and how different images interleaved with text receive different starting offsets. Distinguish an M-RoPE position ID from an index in the flattened token sequence. Why should horizontal and vertical displacement be represented separately, and why should video timing reflect elapsed time rather than only frame order?