Vision Transformers: Oral Knowledge Questions

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

1. Why can a Vision Transformer not determine the original spatial arrangement of image patches from patch content alone? Explain how positional embeddings solve this problem.

2. Explain the roles of queries, keys, and values without simply reciting the attention equation. What determines the attention weights, and what information is actually combined?

3. Inside a Vision Transformer encoder block, how do multi-head self-attention and the MLP perform different jobs? Which dimension does each operation mix?

4. Explain how the final [CLS] token becomes an image-class prediction. In the expression below, what does c mean?

$$p(y = c \mid \text{image}) = \text{softmax}(\text{logits})_c$$

5. If the patch size is reduced from 16 x 16 to 8 x 8, what happens to spatial detail, sequence length, and attention cost? Explain the trade-off rather than giving only the formulas.