

Omni VLMs: Oral Knowledge Questions

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

1. **Components of an image tokenizer.**

What are the three main components of a vector-quantized image tokenizer? Explain the role of each and describe how an input image is converted into discrete code IDs and then reconstructed.

2. From image features to discrete codes.

At each spatial location, the image encoder produces a continuous feature vector. The vector quantizer selects the nearest vector from a learned codebook and records its integer ID.

Explain the difference between the encoder's feature vector, the selected code ID, and the selected codebook vector. Which does the image-generation Transformer predict, and which does the image decoder receive? Why does a code ID such as 7 not represent a pixel brightness?

3. Why is straight-through estimation needed?

Why does nearest-code selection prevent useful ordinary reconstruction gradients from reaching the encoder? Explain how straight-through estimation addresses this problem.

4. **Visible context without target leakage.**

In a MaskGIT training example, the input is

(2, MASK, 1, 8, MASK, MASK).

The correct target at position 2 is code 7.

When predicting position 2, may the model attend to the visible code 8 at position 4? Why does this not reveal the hidden target? How does this differ from the information available when an autoregressive model predicts the second code?

5. **Parallel prediction and iterative generation.**

If MaskGIT predicts all currently masked positions in parallel, why does it use several refinement rounds? Explain the roles of categorical distributions, sampled candidate codes, and confidence scores. How does the procedure change when some image codes must remain fixed for inpainting?