

# SteerViT: Oral Knowledge Questions

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

## 1. **What makes a visual encoder steerable?**

An image contains both a cat and a bookshelf. We first use the prompt “the cat” and then “the bookshelf.”

Explain the difference between changing the text used to score a fixed CLIP image representation and changing the visual representation itself in SteerViT. At what stage does the prompt influence the image processing in each approach?

## 2. How does text enter the visual backbone?

Describe the path from a written prompt to the cross-attention updates inside SteerViT. Include the text encoder, the trainable text adapter, and the roles of image queries and text keys and values.

Why is it useful to retain multiple contextualized prompt-token vectors rather than reduce the entire prompt to one vector?

3. **What does “frozen visual backbone” mean?**

SteerViT freezes the original visual Transformer’s parameters, yet its output changes with the prompt.

Explain how both statements can be true. Distinguish between fixed model weights and fixed intermediate representations. Can gradients pass through frozen visual layers to train an adapter inserted earlier in the network?

#### 4. **Why train with segmentation supervision?**

SteerViT is trained using an image, a referring expression, and a mask identifying the referred object.

Explain how this supervision teaches the model which visual regions should respond to the prompt. Why might supervising the object’s full spatial extent provide a richer training signal than supervising only its center point?

**5. How do we distinguish useful visual steering from merely responding to text?**

A model retrieves the correct object when given an image and an appropriate prompt. Explain why this result alone does not establish that the model is a useful steerable visual encoder.

Describe what each of the following tests would reveal:

- (a) Keep an image containing two objects fixed and switch the prompt from one object to the other. Also test an irrelevant prompt. What would changed retrieval results demonstrate, and why would a performance drop with an irrelevant prompt alone be insufficient?
- (b) Retrieve among visually similar objects using a detailed prompt, such as “the red mug with a white handle.” What capability does this test beyond selecting a broad object category?