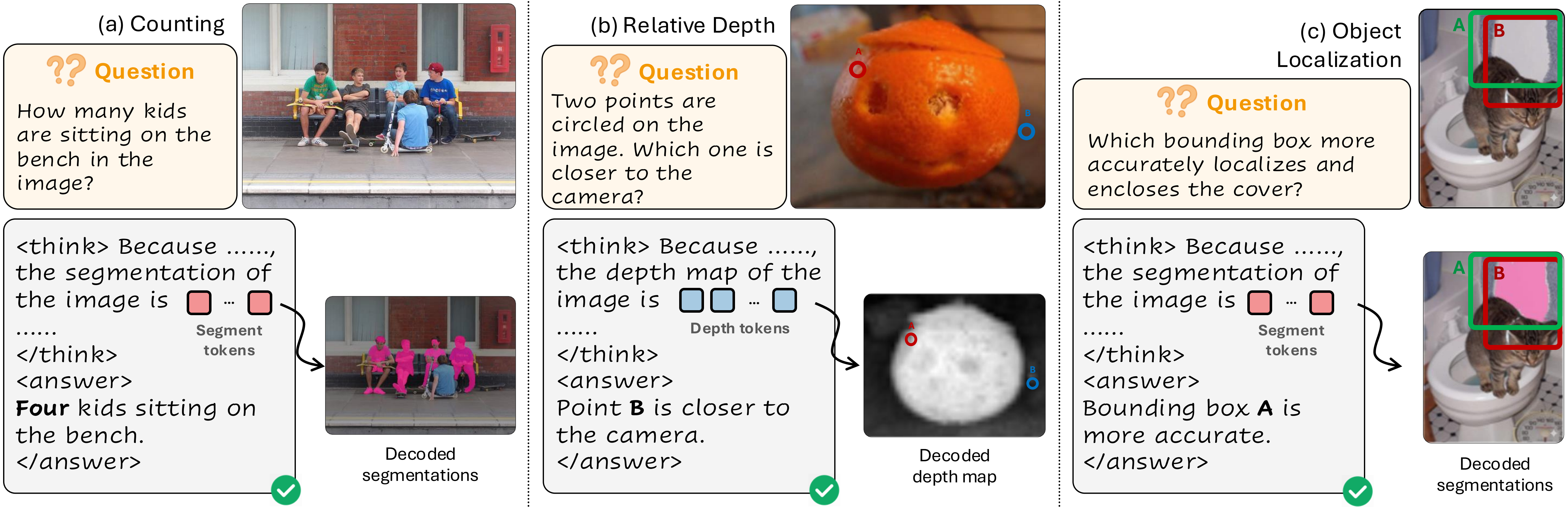

(a) Counting

Question

How many kids are sitting on the bench in the image?

<think> Because ......, the segmentation of the image is 🟥 ... 🟥 ...... </think> <answer> **Four** kids sitting on the bench. </answer> ✓

Segment tokens

Decoded segmentations

(b) Relative Depth

Question

Two points are circled on the image. Which one is closer to the camera?

<think> Because ......, the depth map of the image is 🟦 🟦 ... 🟦 ...... </think> <answer> Point **B** is closer to the camera. </answer> ✓

Depth tokens

Decoded depth map

(c) Object Localization

Question

Which bounding box more accurately localizes and encloses the cover?

<think> Because ......, the segmentation of the image is 🟥 ... 🟥 ...... </think> <answer> Bounding box **A** is more accurate. </answer> ✓

Segment tokens

Decoded segmentations