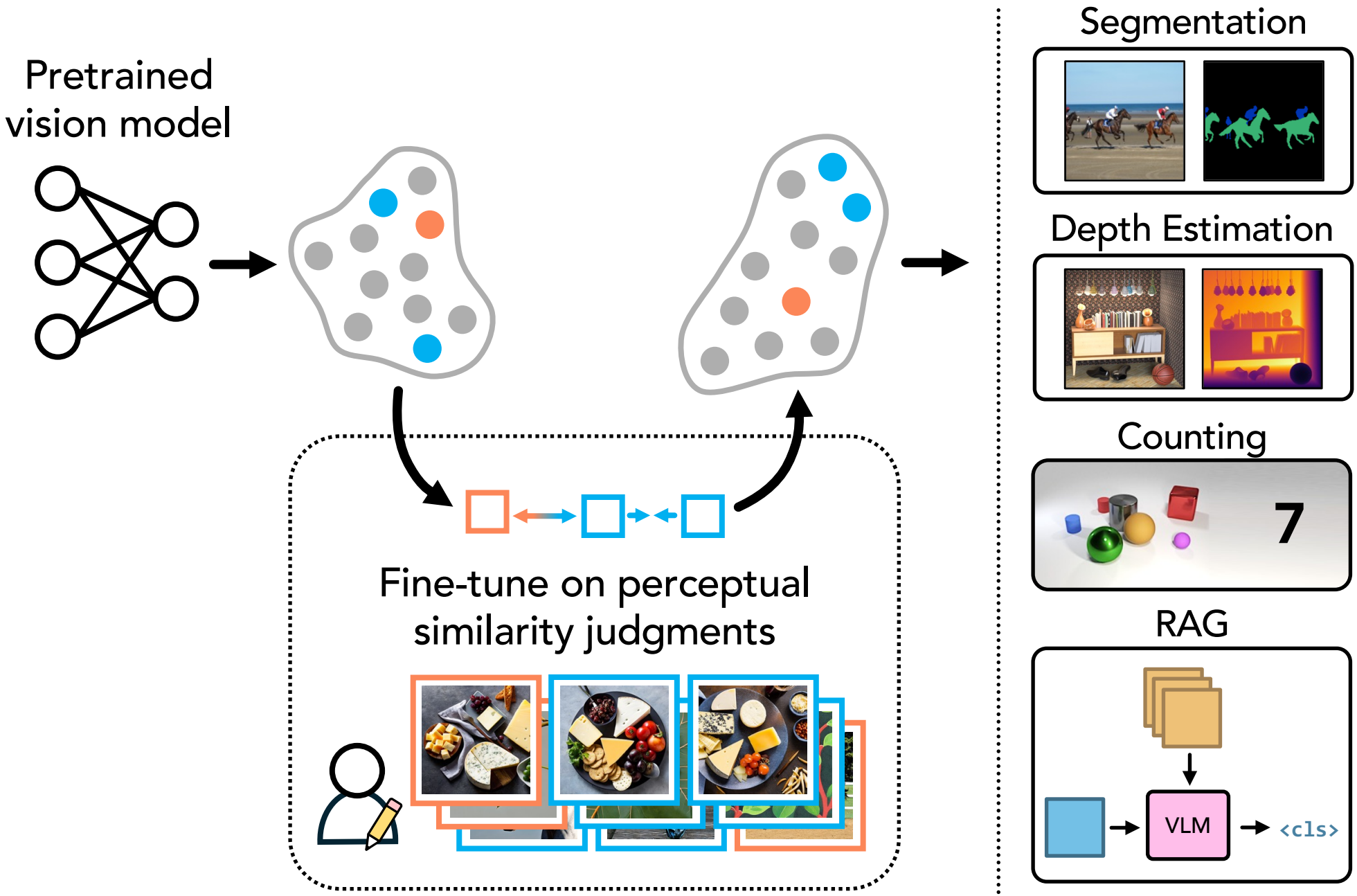
Pretrained vision model
Fine-tune on perceptual similarity judgments
Segmentation
Depth Estimation
Counting
7
RAG
VLM
<cls>