Table 1: Core Vision Tasks

Task Type	Description	Strong VLM Models
Image Classification	Predicts a label for the whole image (e.g., bear, dog, car...)	**CLIP**, **BLIP**, **ALIGN**, **Florence**
Object Detection	Detects multiple objects with bounding boxes	**Grounding DINO**, **OWL-ViT**, **GLIP**
Semantic Segmentation	Labels each pixel by class (not separating instances)	**CLIPSeg**, **LSeg**
Instance Segmentation	Labels and distinguishes each object instance individually	**SAM**, **SEEM**, **GRIT**