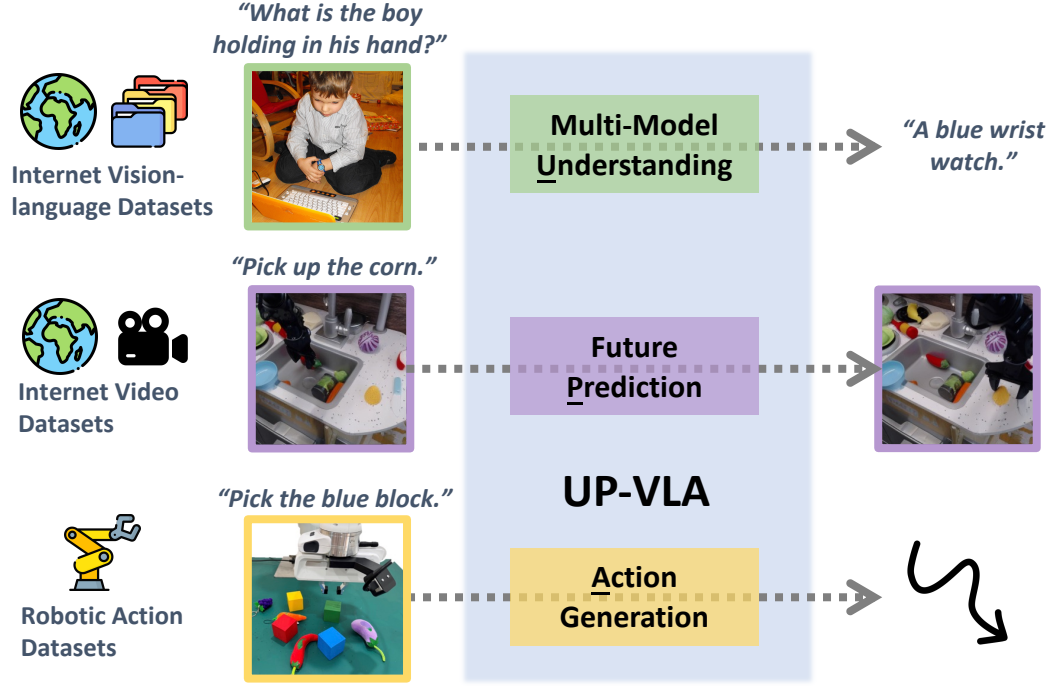
"What is the boy holding in his hand?"
Internet Vision-language Datasets
Multi-Model Understanding
"A blue wrist watch."
"Pick up the corn."
Internet Video Datasets
Future Prediction
UP-VLA
"Pick the blue block."
Robotic Action Datasets
Action Generation