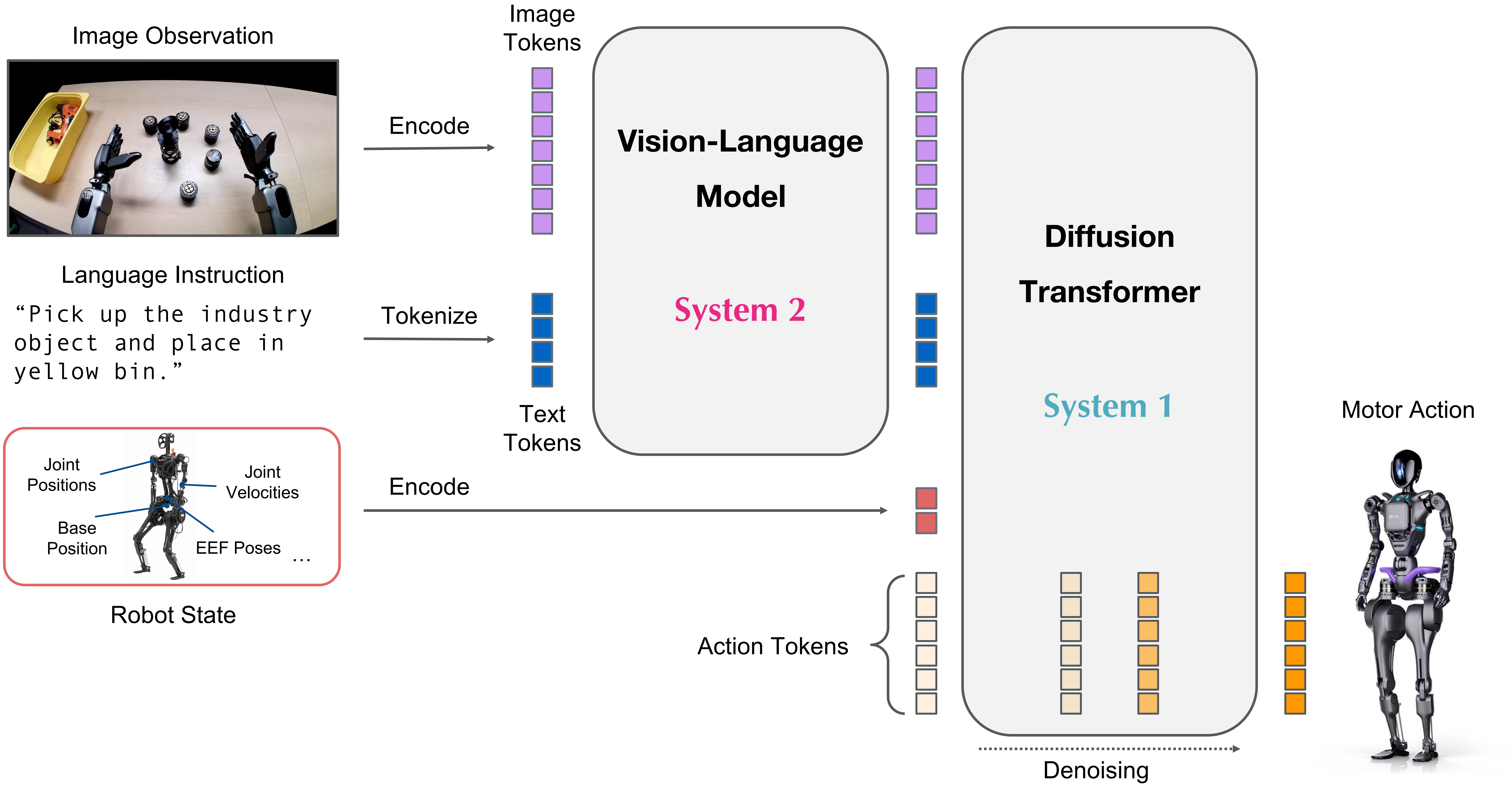
Image Observation
Encode
Image Tokens
Language Instruction
“Pick up the industry object and place in yellow bin.”
Tokenize
Text Tokens
Joint Positions
Joint Velocities
Base Position
EEF Poses
…
Robot State
Encode
Vision-Language Model
System 2
Diffusion Transformer
System 1
Action Tokens
Denoising
Motor Action