

LangToMo

Image x_i

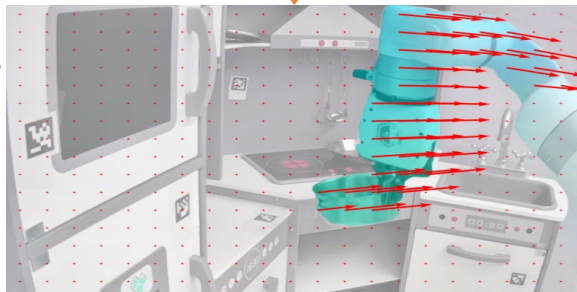


Image x_{i+j}



Language Instruction c
"Place dish in sink"

Pixel Motion
Prediction y_i



System 2 @ 0.5 FPS
(Motion Generation)

System 1 @ 5 FPS
(Motion to Action Mapping)

Visualized Action a_{i+j}



Action Vector a_{i+j}