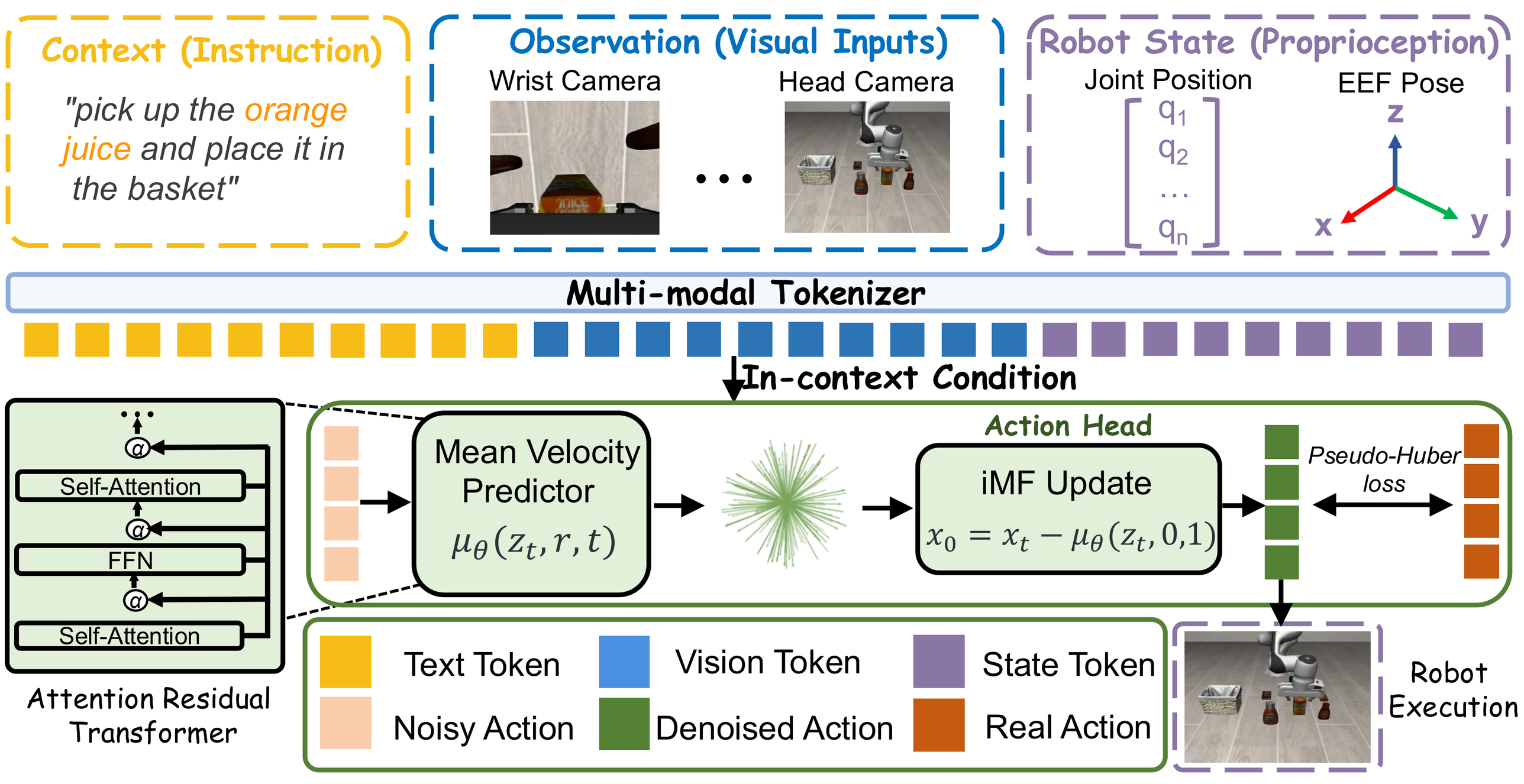
Context (Instruction)
"pick up the orange juice and place it in the basket"
Observation (Visual Inputs)
Wrist Camera
Head Camera
Robot State (Proprioception)
Joint Position
EEF Pose
Multi-modal Tokenizer
In-context Condition
Action Head
Mean Velocity Predictor
$\mu_\theta(z_t, r, t)$
iMF Update
$x_0 = x_t - \mu_\theta(z_t, 0, 1)$
Pseudo-Huber loss
Self-Attention
FFN
Self-Attention
Attention Residual Transformer
Text Token
Vision Token
State Token
Noisy Action
Denoised Action
Real Action
Robot Execution