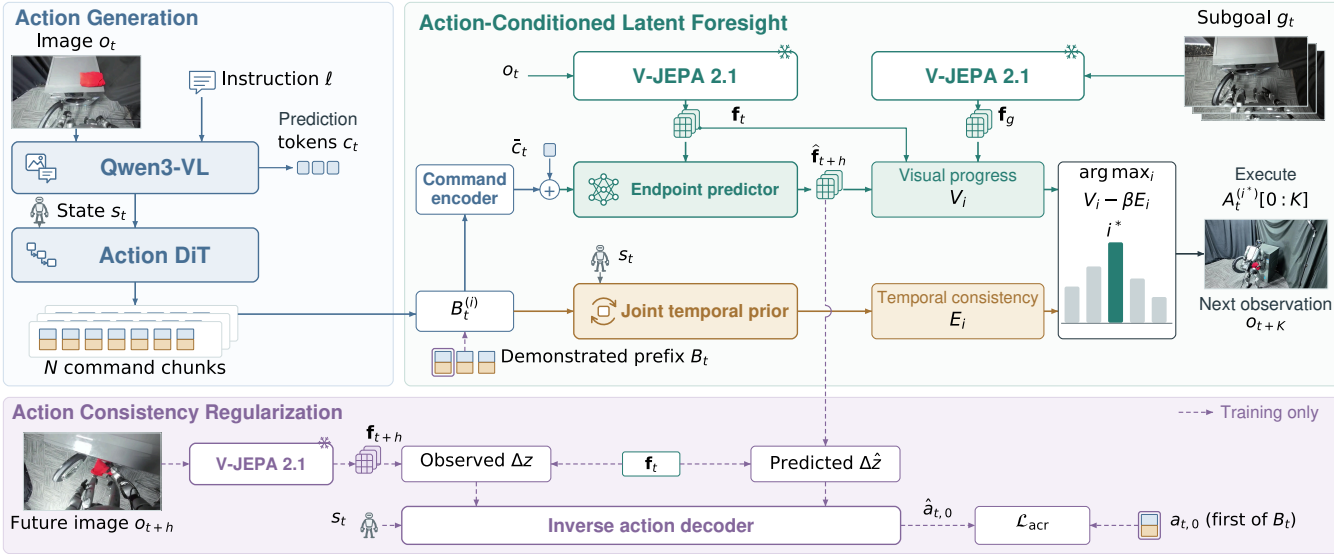

Action Generation
Image o_t
Instruction ℓ
Prediction tokens c_t
Qwen3-VL
State s_t
Action DiT
N command chunks
Action-Conditioned Latent Foresight
Subgoal g_t
o_t
V-JEPA 2.1
V-JEPA 2.1
$\mathbf{f}_t$
$\mathbf{f}_g$
$\bar{c}_t$
Command encoder
Endpoint predictor
$\hat{\mathbf{f}}_{t+h}$
Visual progress V_i
$\arg\max_i$ $V_i - \beta E_i$
i^*
Execute $A_t^{(i^*)}[0:K]$
Next observation o_{t+K}
s_t
$B_t^{(i)}$
Joint temporal prior
Temporal consistency E_i
Demonstrated prefix B_t
Action Consistency Regularization
Training only
Future image o_{t+h}
V-JEPA 2.1
$\mathbf{f}_{t+h}$
Observed Δz
$\mathbf{f}_t$
Predicted $\Delta\hat{z}$
s_t
Inverse action decoder
$\hat{a}_{t,0}$
$\mathcal{L}_{\text{acr}}$
$a_{t,0}$ (first of B_t)