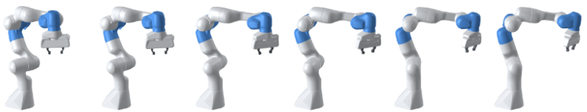
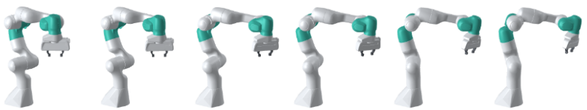
Physical action neighborhoods

$$d_{phys}(a,i) = \alpha_p {\delta_p}^2 + \alpha_R {\delta_R}^2 + \alpha_g {\delta_g}^2$$

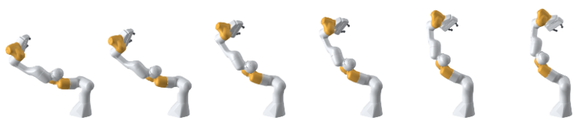
Anchor Action a



Near Action n



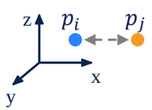
Far Action f



$$d_{\text{phys}}(a,n) < d_{\text{phys}}(a,f)$$

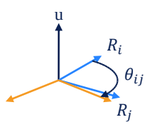
Physical distance components

Translation



$${\delta_p}^2 = \|\Delta p / s_p\|^2$$

Rotation

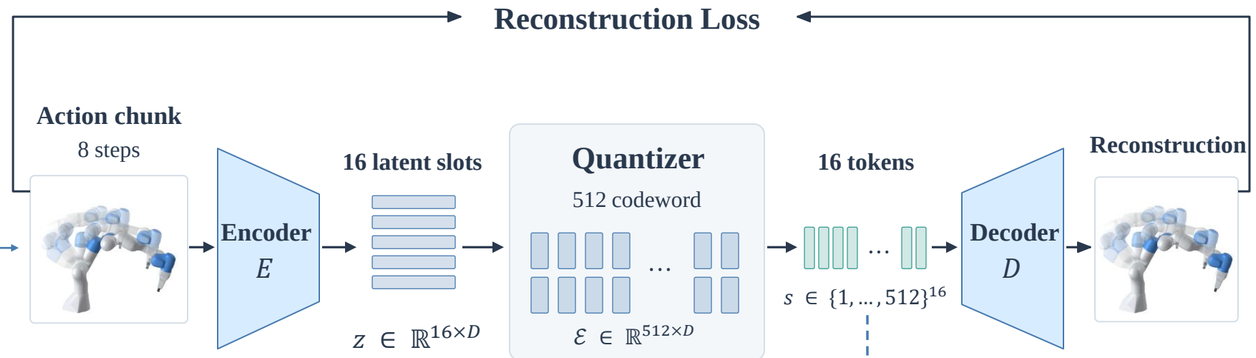


$${\delta_R}^2 = (\theta / s_R)^2$$

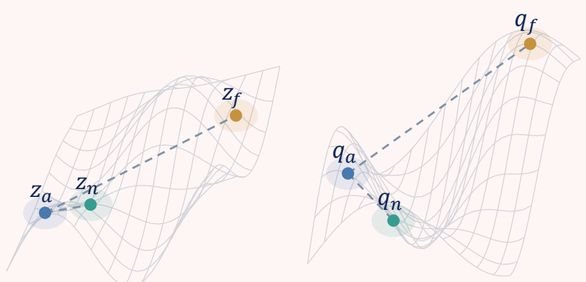
Gripper

$${\delta_g}^2 = |\Delta g|^2$$

Action tokenizer



PRP Loss: Physical Rank Preservation



Encoder features z — $L_2(z_a, z_i)$

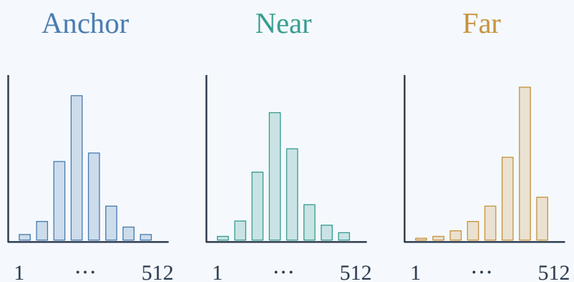
Quantized features q — $L_2(q_a, q_i)$

$$d_{ij}{}^2 = 0.25 \cdot L_2(z_i, z_j) + 0.75 \cdot L_2(q_i, q_j)$$

$$d(anchor, near) < d(anchor, far)$$

QR Loss: Quantization Regularization

Codeword assignment distributions



$$JS(P,Q) = \frac{1}{2} KL(P\|m) + \frac{1}{2} KL(Q\|m), \quad m = \frac{P+Q}{2}$$

$$JS(anchor, near) < JS(anchor, far)$$

Policy learning

Image + instruction

Image

Instruction

Grab the red can and put it in the box.

Autoregressive VLA
Next-token prediction

16 action tokens

$s \in \{1,\dots,512\}^{16}$

Frozen decoder
D

Robot actions
8-step action chunk