$$a_{t:t+H_a}$$

VLM Backbone

Action Expert
$$p_\theta(a_{t:t+H_a}|o_t)$$

$$\mathcal{T}_i$$

$$\mathcal{T}_q$$

$$\mathcal{T}_a$$

Image Emb n

Image Emb i

Image Emb 0

Language Embedding

"Pick up the box"

q Emb

a Emb

$$q_t: \begin{matrix} -1.3628, \\ 78.2137, \\ -97.150, \\ 96.1393, \\ 18.8767, \\ 44.9446 \end{matrix}$$

$$\tilde{a}_{t:t+H_a}$$