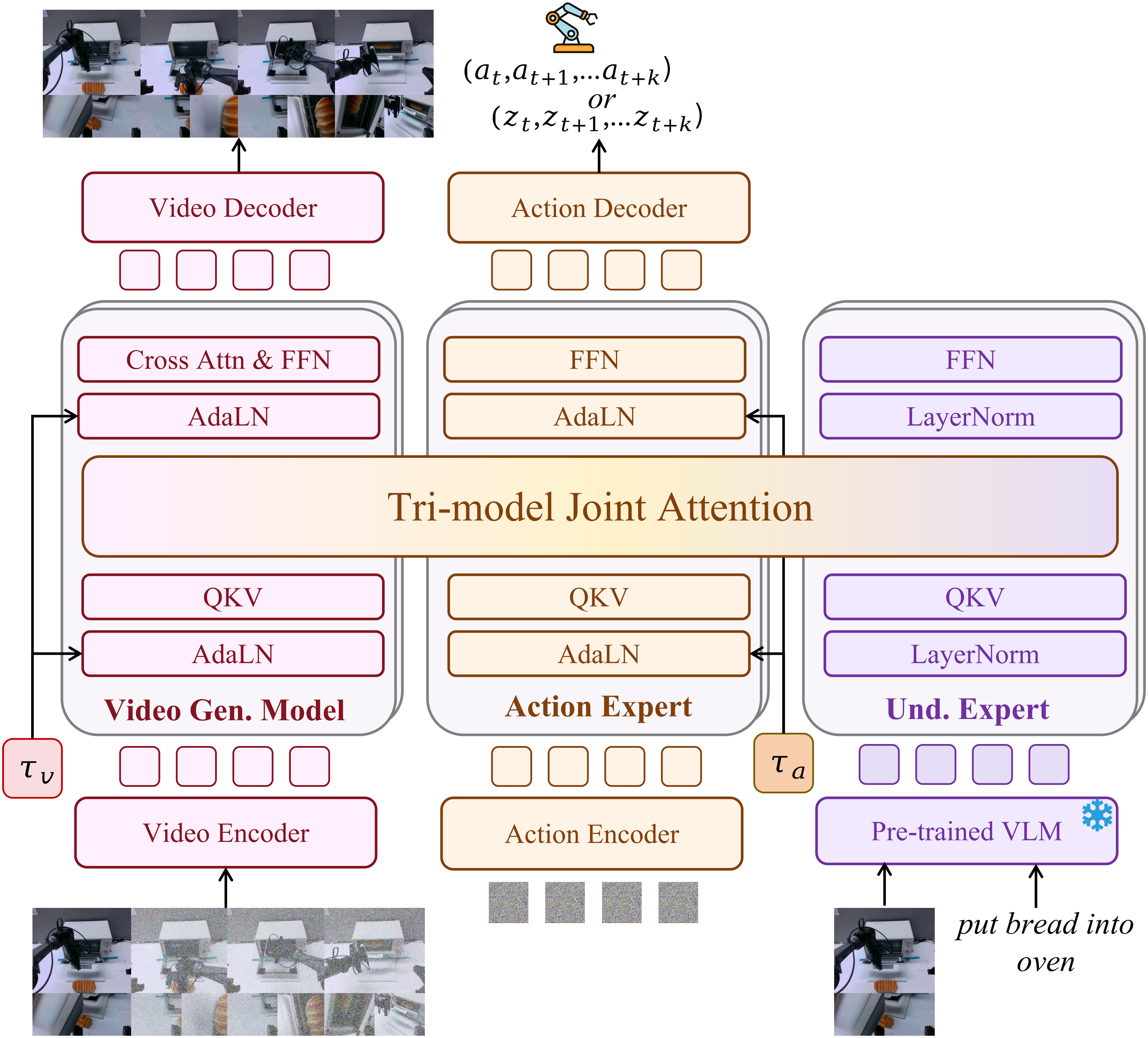

$(a_t, a_{t+1}, \ldots a_{t+k})$
or
$(z_t, z_{t+1}, \ldots z_{t+k})$
Video Decoder
Action Decoder
Cross Attn & FFN
AdaLN
FFN
AdaLN
FFN
LayerNorm
Tri-model Joint Attention
QKV
AdaLN
QKV
AdaLN
QKV
LayerNorm
Video Gen. Model
Action Expert
Und. Expert
τ_v
τ_a
Video Encoder
Action Encoder
Pre-trained VLM
put bread into oven