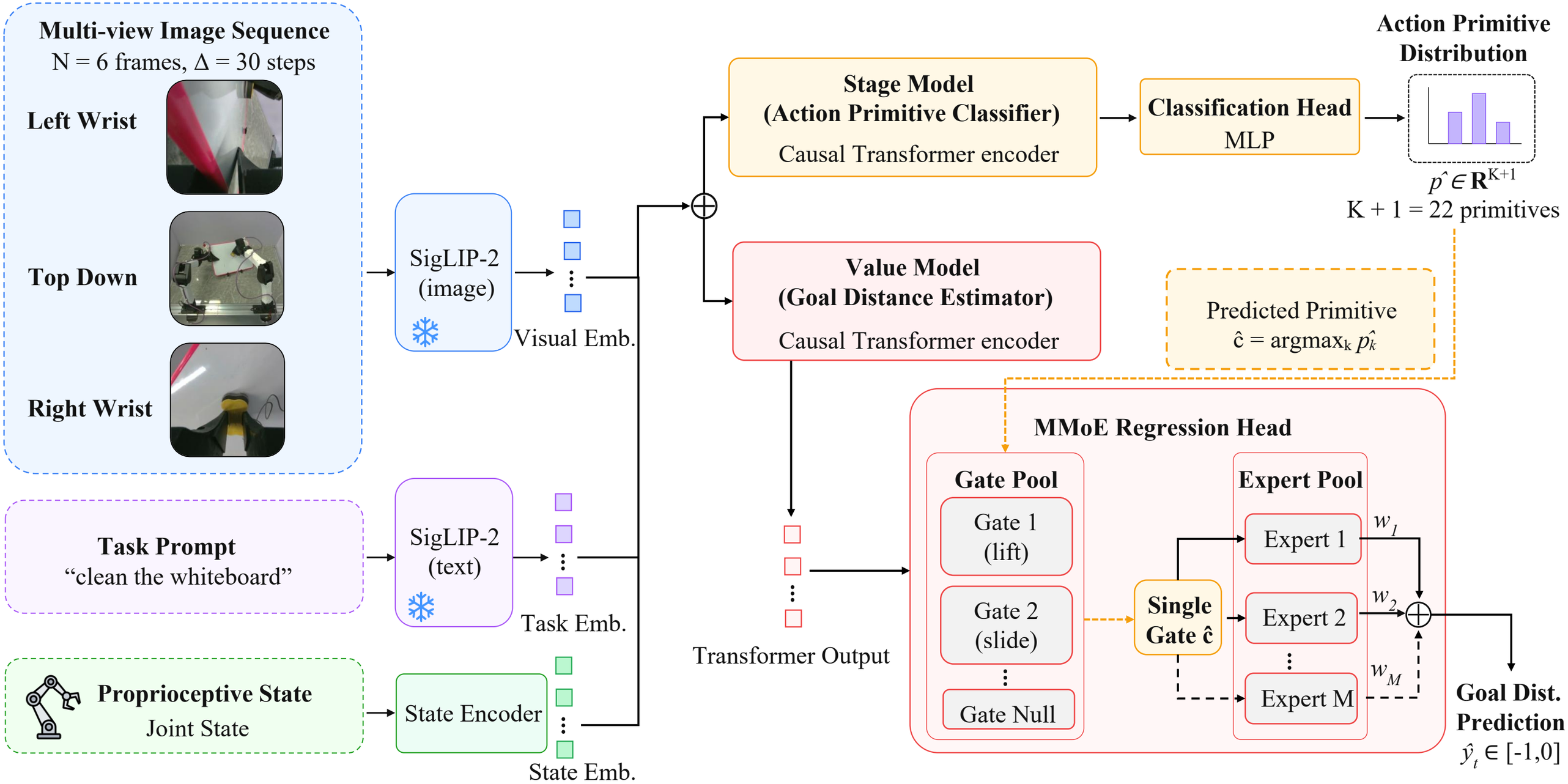
Multi-view Image Sequence
N = 6 frames, Δ = 30 steps
Left Wrist
Top Down
Right Wrist
Task Prompt
“clean the whiteboard”
Proprioceptive State
Joint State
SigLIP-2 (image)
SigLIP-2 (text)
State Encoder
Visual Emb.
Task Emb.
State Emb.
Stage Model (Action Primitive Classifier)
Causal Transformer encoder
Classification Head
MLP
Action Primitive Distribution
$\hat{p} \in \mathbf{R}^{K+1}$
K + 1 = 22 primitives
Value Model (Goal Distance Estimator)
Causal Transformer encoder
Transformer Output
Predicted Primitive
$\hat{c} = \text{argmax}_k \hat{p}_k$
MMoE Regression Head
Gate Pool
Gate 1 (lift)
Gate 2 (slide)
Gate Null
Single Gate $\hat{c}$
Expert Pool
Expert 1
Expert 2
Expert M
w_1
w_2
w_M
Goal Dist. Prediction
$\hat{y}_t \in [-1,0]$