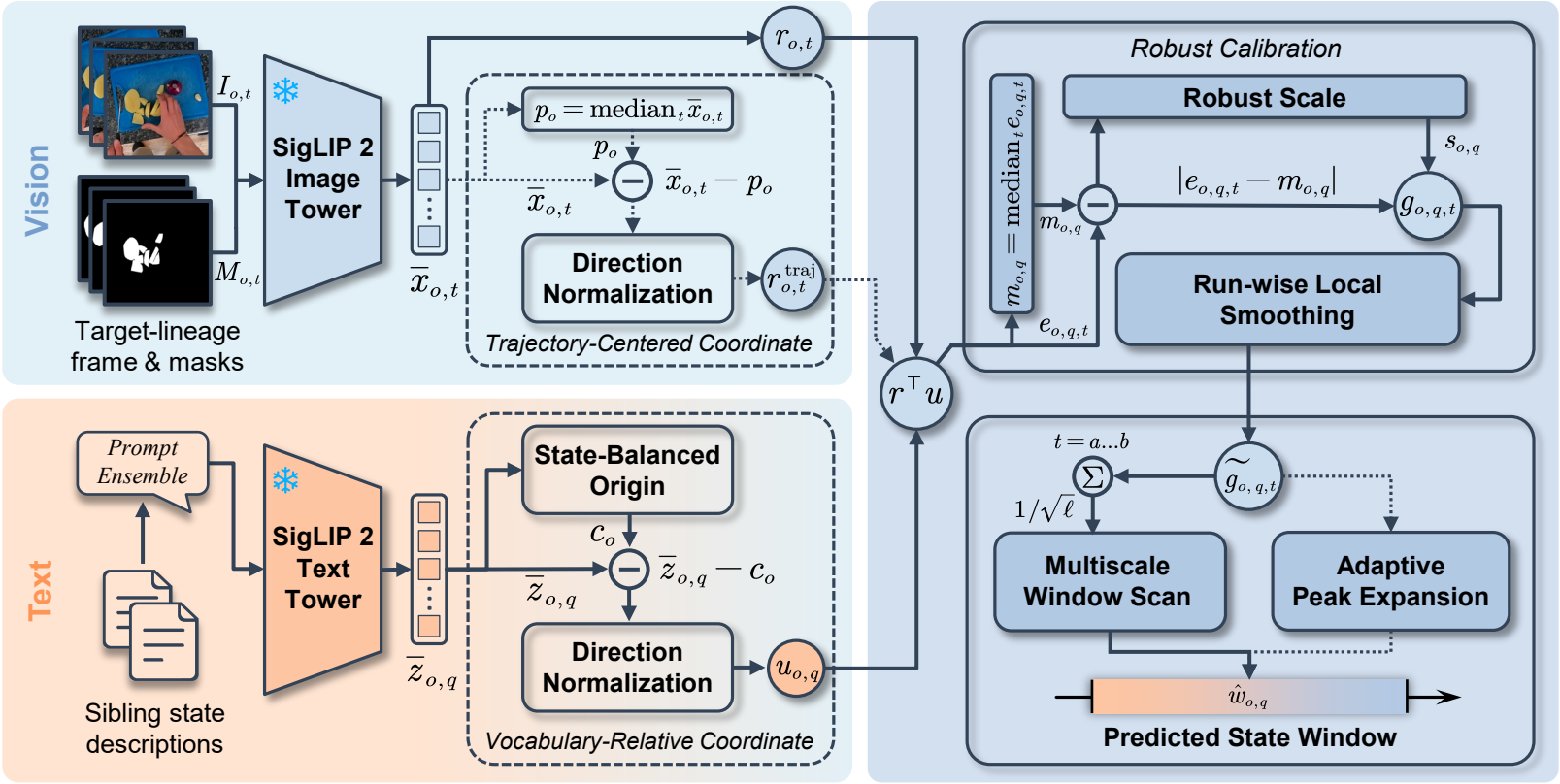

Vision
$I_{o,t}$
$M_{o,t}$
SigLIP 2 Image Tower
$\bar{x}_{o,t}$
Target-lineage frame & masks
$r_{o,t}$
$p_o = \text{median}_t\, \bar{x}_{o,t}$
p_o
$\bar{x}_{o,t}$
$\bar{x}_{o,t} - p_o$
Direction Normalization
$r^{\text{traj}}_{o,t}$
Trajectory-Centered Coordinate
Text
Prompt Ensemble
Sibling state descriptions
SigLIP 2 Text Tower
$\bar{z}_{o,q}$
State-Balanced Origin
c_o
$\bar{z}_{o,q}$
$\bar{z}_{o,q} - c_o$
Direction Normalization
$u_{o,q}$
Vocabulary-Relative Coordinate
$r^\top u$
Robust Calibration
$m_{o,q} = \text{median}_t\, e_{o,q,t}$
$m_{o,q}$
$e_{o,q,t}$
Robust Scale
$s_{o,q}$
$|e_{o,q,t} - m_{o,q}|$
$g_{o,q,t}$
Run-wise Local Smoothing
$\tilde{g}_{o,q,t}$
$t = a...b$
Σ
$1/\sqrt{\ell}$
Multiscale Window Scan
Adaptive Peak Expansion
$\hat{w}_{o,q}$
Predicted State Window