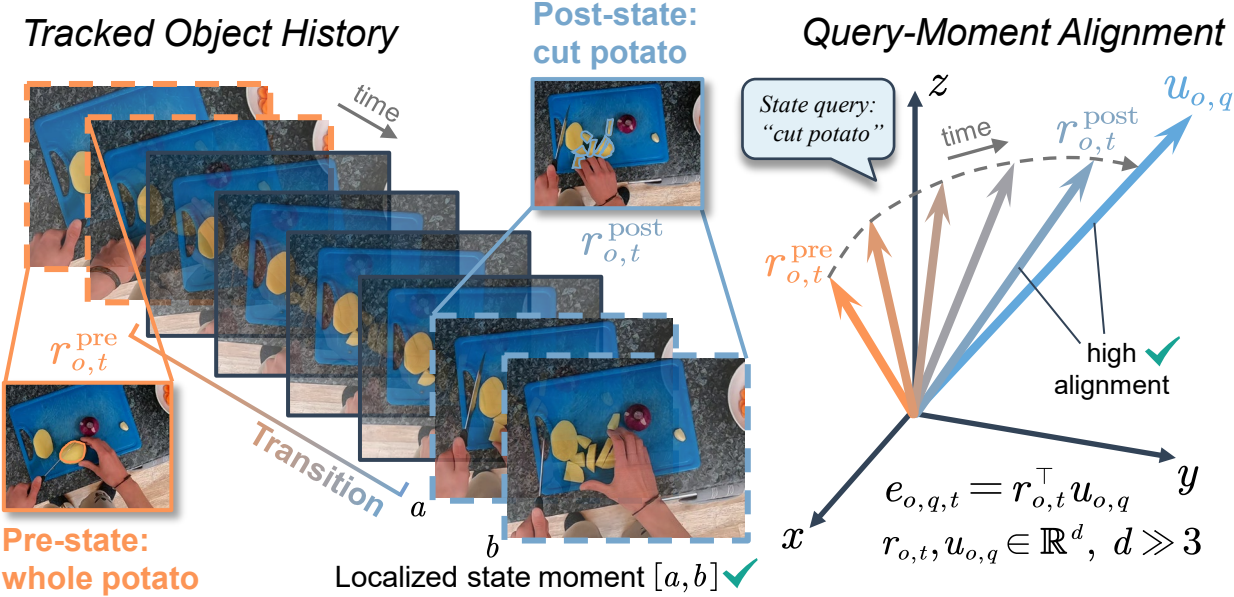
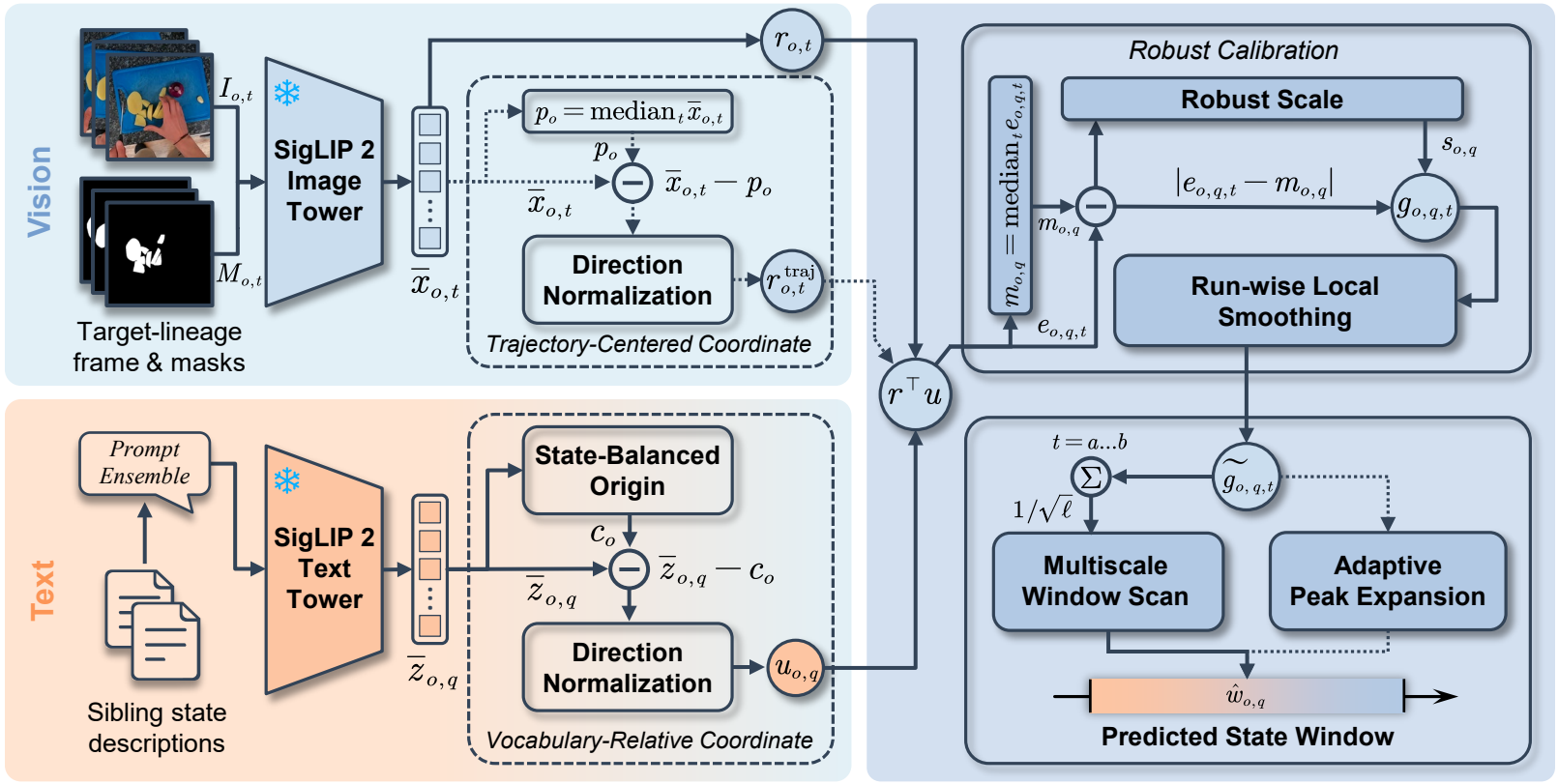

Tracked Object History
Post-state:
cut potato
Query-Moment Alignment
time
$r^{\text{post}}_{o,t}$
$r^{\text{pre}}_{o,t}$
Transition
a
b
Pre-state:
whole potato
Localized state moment $[a, b]$
State query:
"cut potato"
z
time
$u_{o,q}$
$r^{\text{post}}_{o,t}$
$r^{\text{pre}}_{o,t}$
high
alignment
y
x
$e_{o,q,t} = r^{\top}_{o,t} u_{o,q}$
$r_{o,t}, u_{o,q} \in \mathbb{R}^d,\ d \gg 3$


Vision
Target-lineage frame & masks
SigLIP 2 Image Tower
Trajectory-Centered Coordinate
Direction Normalization
Text
Prompt Ensemble
Sibling state descriptions
SigLIP 2 Text Tower
State-Balanced Origin
Vocabulary-Relative Coordinate
Direction Normalization
Robust Calibration
Robust Scale
Run-wise Local Smoothing
Multiscale Window Scan
Adaptive Peak Expansion
Predicted State Window