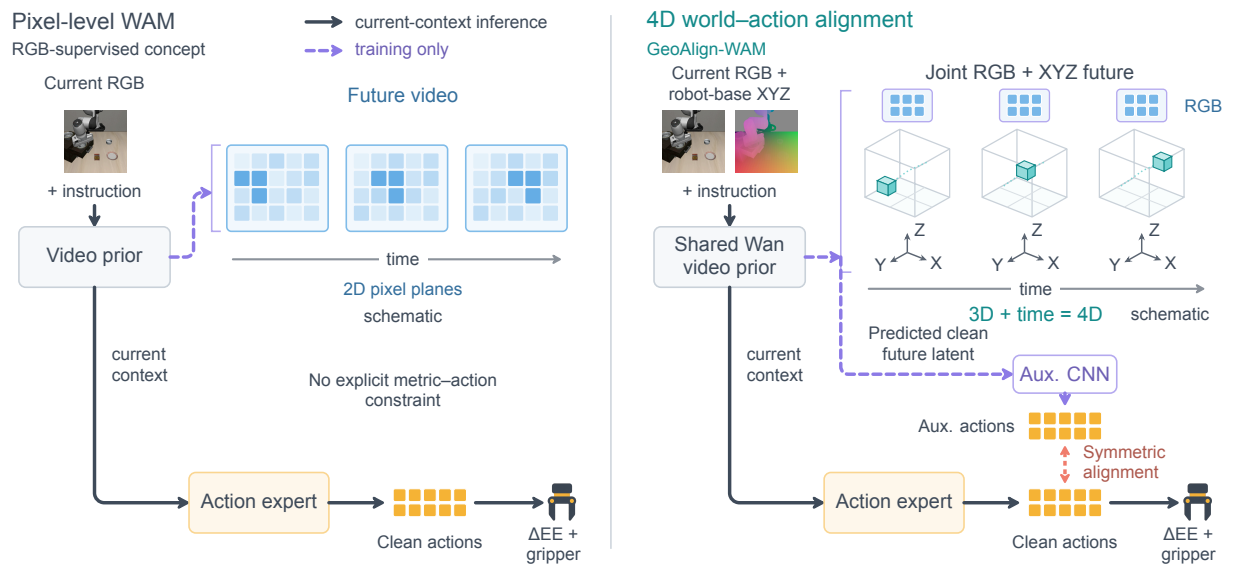

Pixel-level WAM
RGB-supervised concept
current-context inference
training only
Current RGB
+ instruction
Video prior
Future video
time
2D pixel planes
schematic
current context
No explicit metric–action constraint
Action expert
Clean actions
ΔEE + gripper
4D world–action alignment
GeoAlign-WAM
Current RGB + robot-base XYZ
+ instruction
Shared Wan video prior
Joint RGB + XYZ future
RGB
Z
Y
X
time
3D + time = 4D
schematic
Predicted clean future latent
current context
Aux. CNN
Aux. actions
Symmetric alignment
Action expert
Clean actions
ΔEE + gripper