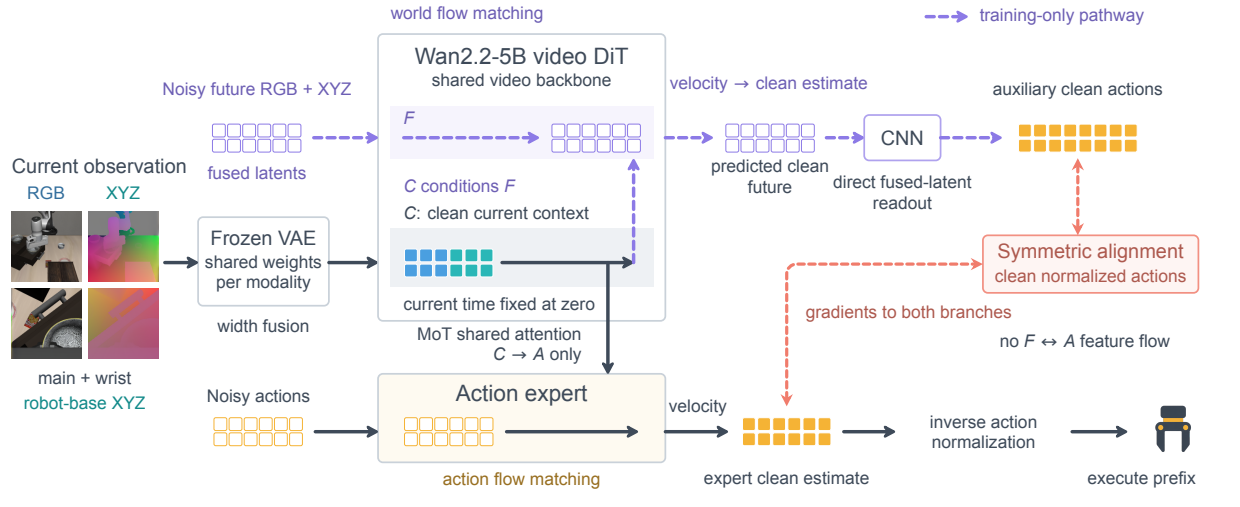

world flow matching
training-only pathway
Wan2.2-5B video DiT
shared video backbone
Noisy future RGB + XYZ
velocity → clean estimate
auxiliary clean actions
F
fused latents
predicted clean future
CNN
direct fused-latent readout
Current observation
RGB
XYZ
C conditions F
C: clean current context
Frozen VAE
shared weights per modality
current time fixed at zero
Symmetric alignment
clean normalized actions
gradients to both branches
no F ↔ A feature flow
width fusion
MoT shared attention
C → A only
main + wrist
robot-base XYZ
Noisy actions
Action expert
velocity
inverse action normalization
action flow matching
expert clean estimate
execute prefix