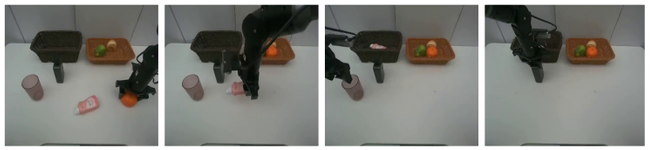

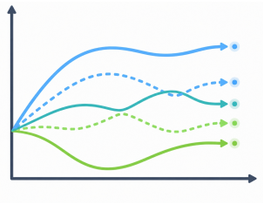

Cross Attention

Residual Update

Slow

Video DiT
(World Planner)

Fast

Action DiT
(Action Expert)

× L

Cross Attention & FFN

AdaLN

Layer-wise
Joint Attention

Video Context Routing

Self Attntion

K / V

Joint Attention

Q K V

AdaLN

Obs-guided query

VAE Encoder

Vision Encoder

State Encoder

Rolling K/V Memory

Video Branch

Action Chunk1

Action Chunk2

o^1

o^2

f_1 f_2 f_n a_1^1 a_2^1 a_1^2 a_2^2

Training

f_1 a_1^1 a_2^1 a_1^2 a_2^2

Inference

Obs-guided query

Updated K / V