

Stage 1 Action Representation Pretraining

Human Manipulation Data



Egocentric Video Clips

Hand Pose Actions

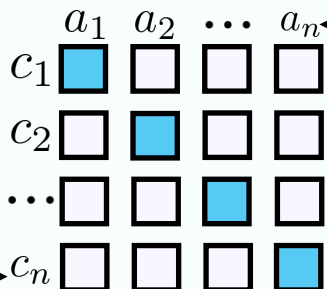
 Frozen
V-JEPA 2.1

Action Encoder

MAP + Proj

MAP + Proj

Noise-Augmented
Latent Decoding



Sigmoid Contrastive Loss

Stage 1b

Action Decoder

Action Reconstruction
Loss

Stage 2 Generation in the Pretrained Action Space

Target Action Chunk



 Frozen

Action Encoder

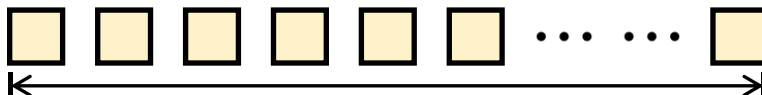
Noisy Action Representation



DiT

 Frozen

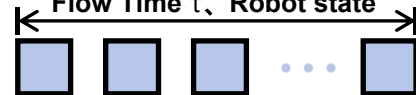
Action Decoder



Action Chunk

Conditions:

Flow Time τ , Robot state



Vision Observations

 Frozen

V-JEPA 2.1

Cross-Attention

Adaptive RMSNorm