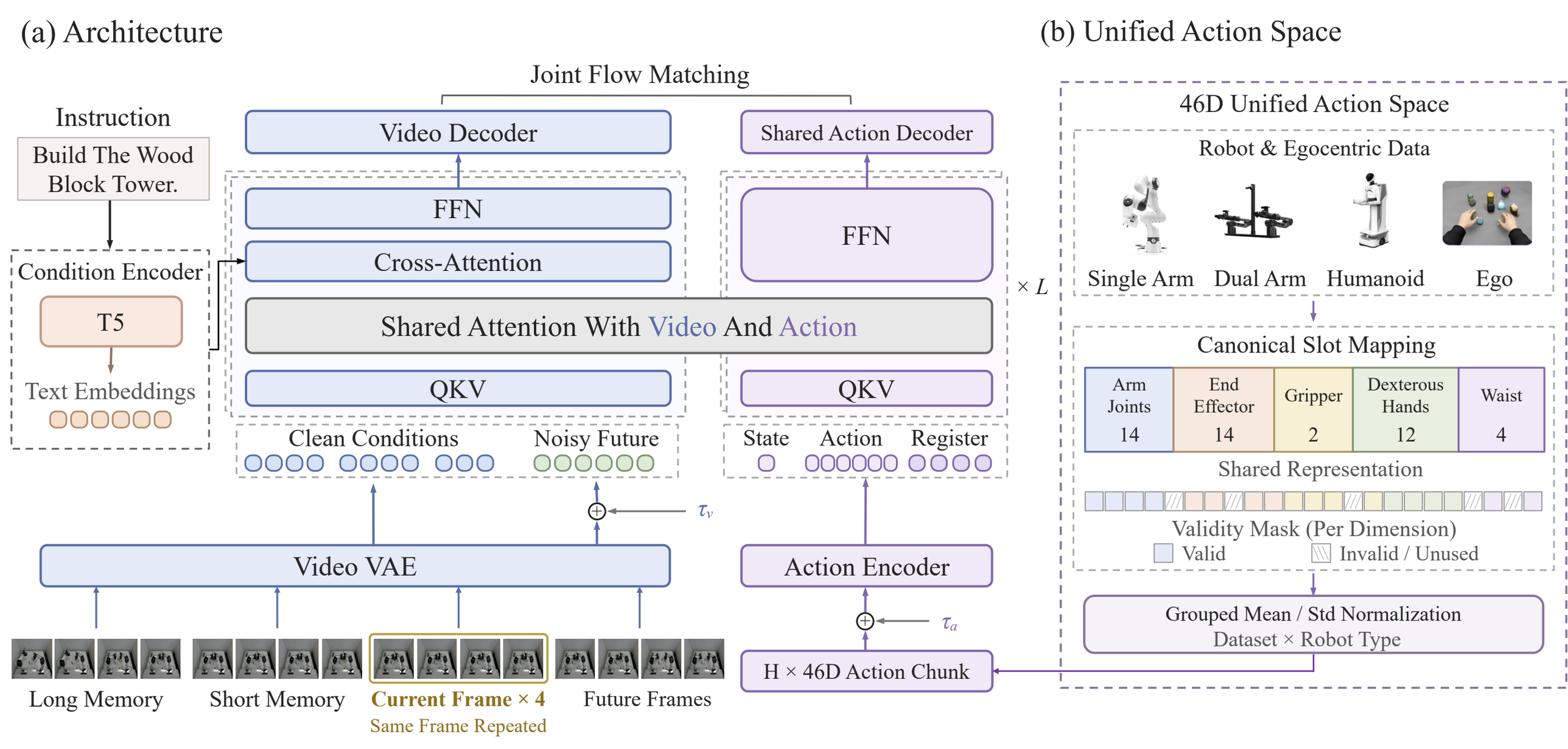

(a) Architecture
Joint Flow Matching
Instruction
Build The Wood Block Tower.
Condition Encoder
T5
Text Embeddings
Video Decoder
Shared Action Decoder
FFN
Cross-Attention
FFN
Shared Attention With Video And Action
QKV
QKV
× L
Clean Conditions
Noisy Future
State
Action
Register
τ_v
τ_a
Video VAE
Action Encoder
H × 46D Action Chunk
Long Memory
Short Memory
Current Frame × 4
Same Frame Repeated
Future Frames
(b) Unified Action Space
46D Unified Action Space
Robot & Egocentric Data
Single Arm
Dual Arm
Humanoid
Ego
Canonical Slot Mapping
Arm Joints 14
End Effector 14
Gripper 2
Dexterous Hands 12
Waist 4
Shared Representation
Validity Mask (Per Dimension)
Valid
Invalid / Unused
Grouped Mean / Std Normalization
Dataset × Robot Type