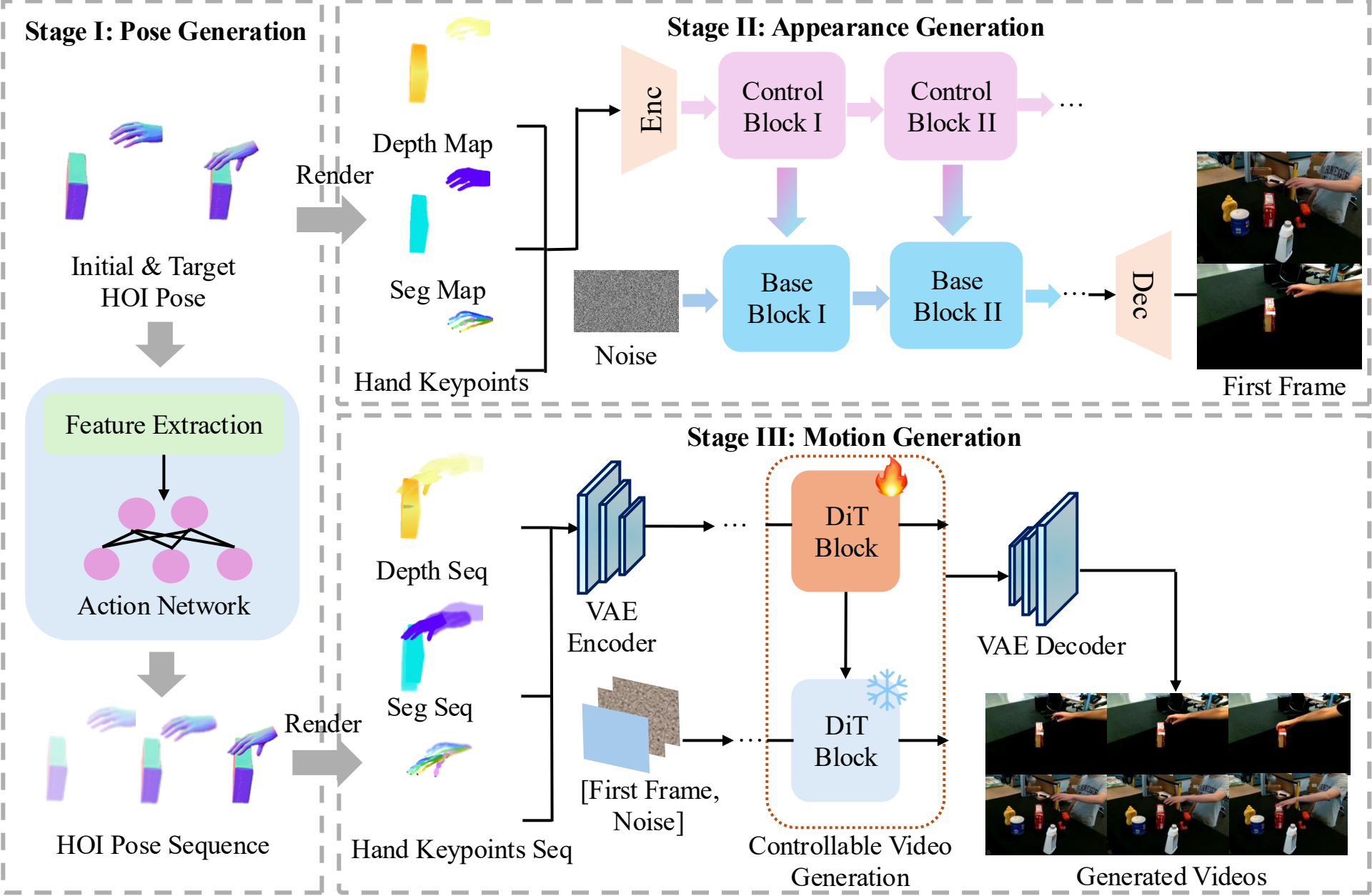
Stage I: Pose Generation
Initial & Target HOI Pose
Feature Extraction
Action Network
HOI Pose Sequence
Render
Depth Map
Seg Map
Hand Keypoints
Stage II: Appearance Generation
Enc
Control Block I
Control Block II
Base Block I
Base Block II
Noise
Dec
First Frame
Stage III: Motion Generation
Depth Seq
Seg Seq
Hand Keypoints Seq
VAE Encoder
[First Frame, Noise]
DiT Block
DiT Block
Controllable Video Generation
VAE Decoder
Generated Videos