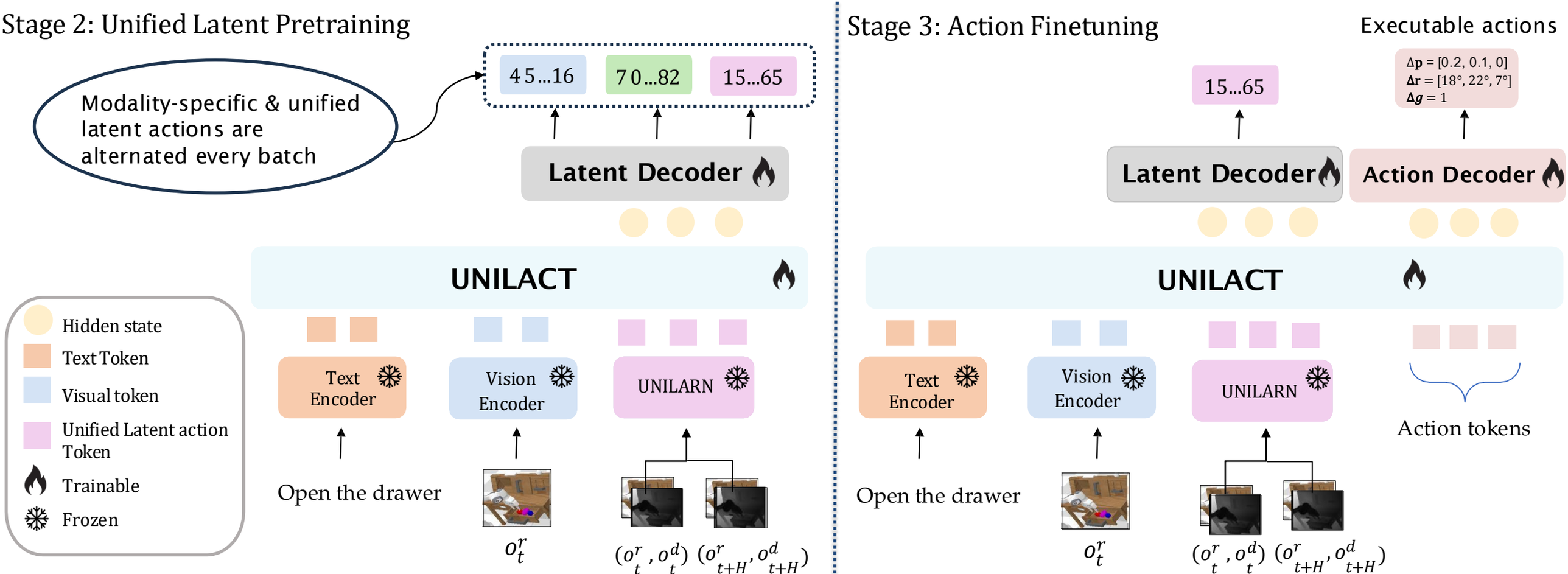
Stage 2: Unified Latent Pretraining
Modality-specific & unified latent actions are alternated every batch
45...16
70...82
15...65
Latent Decoder
UNILACT
Text Encoder
Vision Encoder
UNILARN
Open the drawer
o_t^r
(o_t^r, o_t^d) (o_{t+H}^r, o_{t+H}^d)
Hidden state
Text Token
Visual token
Unified Latent action Token
Trainable
Frozen
Stage 3: Action Finetuning
Executable actions
15...65
Δp = [0.2, 0.1, 0]
Δr = [18°, 22°, 7°]
Δg = 1
Latent Decoder
Action Decoder
UNILACT
Text Encoder
Vision Encoder
UNILARN
Action tokens
Open the drawer
o_t^r
(o_t^r, o_t^d) (o_{t+H}^r, o_{t+H}^d)