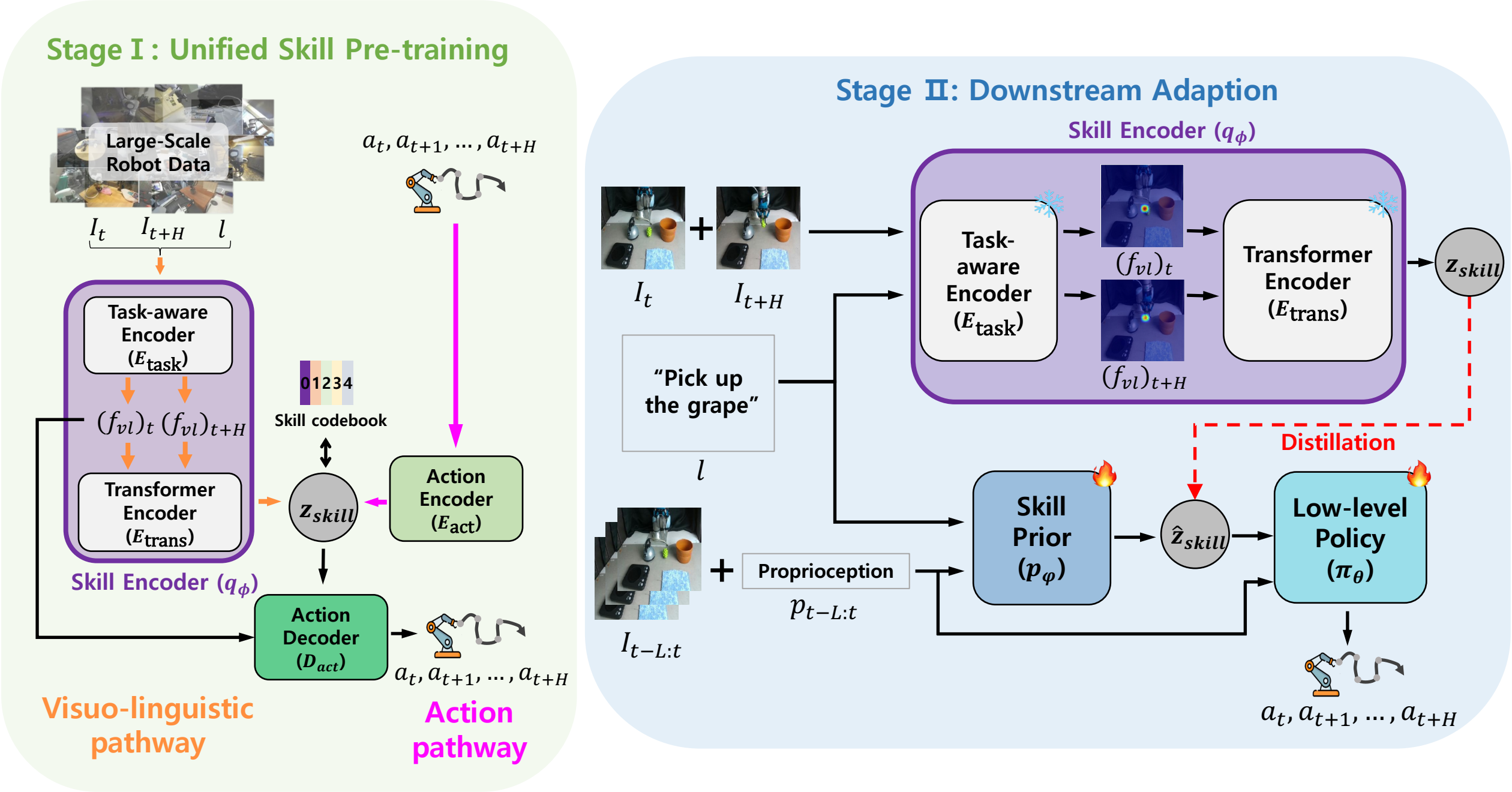
Stage I : Unified Skill Pre-training
Large-Scale Robot Data
I_t I_{t+H} l
$a_t, a_{t+1}, \dots, a_{t+H}$
Task-aware Encoder (E_{task})
$(f_{vl})_t$ $(f_{vl})_{t+H}$
Transformer Encoder (E_{trans})
Skill Encoder (q_ϕ)
01234
Skill codebook
z_{skill}
Action Encoder (E_{act})
Action Decoder (D_{act})
$a_t, a_{t+1}, \dots, a_{t+H}$
Visuo-linguistic pathway
Action pathway
Stage II: Downstream Adaption
Skill Encoder (q_ϕ)
I_t I_{t+H}
Task-aware Encoder (E_{task})
$(f_{vl})_t$
$(f_{vl})_{t+H}$
Transformer Encoder (E_{trans})
z_{skill}
"Pick up the grape"
l
Distillation
$I_{t-L:t}$
Proprioception
$p_{t-L:t}$
Skill Prior (p_φ)
$\hat{z}_{skill}$
Low-level Policy (π_θ)
$a_t, a_{t+1}, \dots, a_{t+H}$