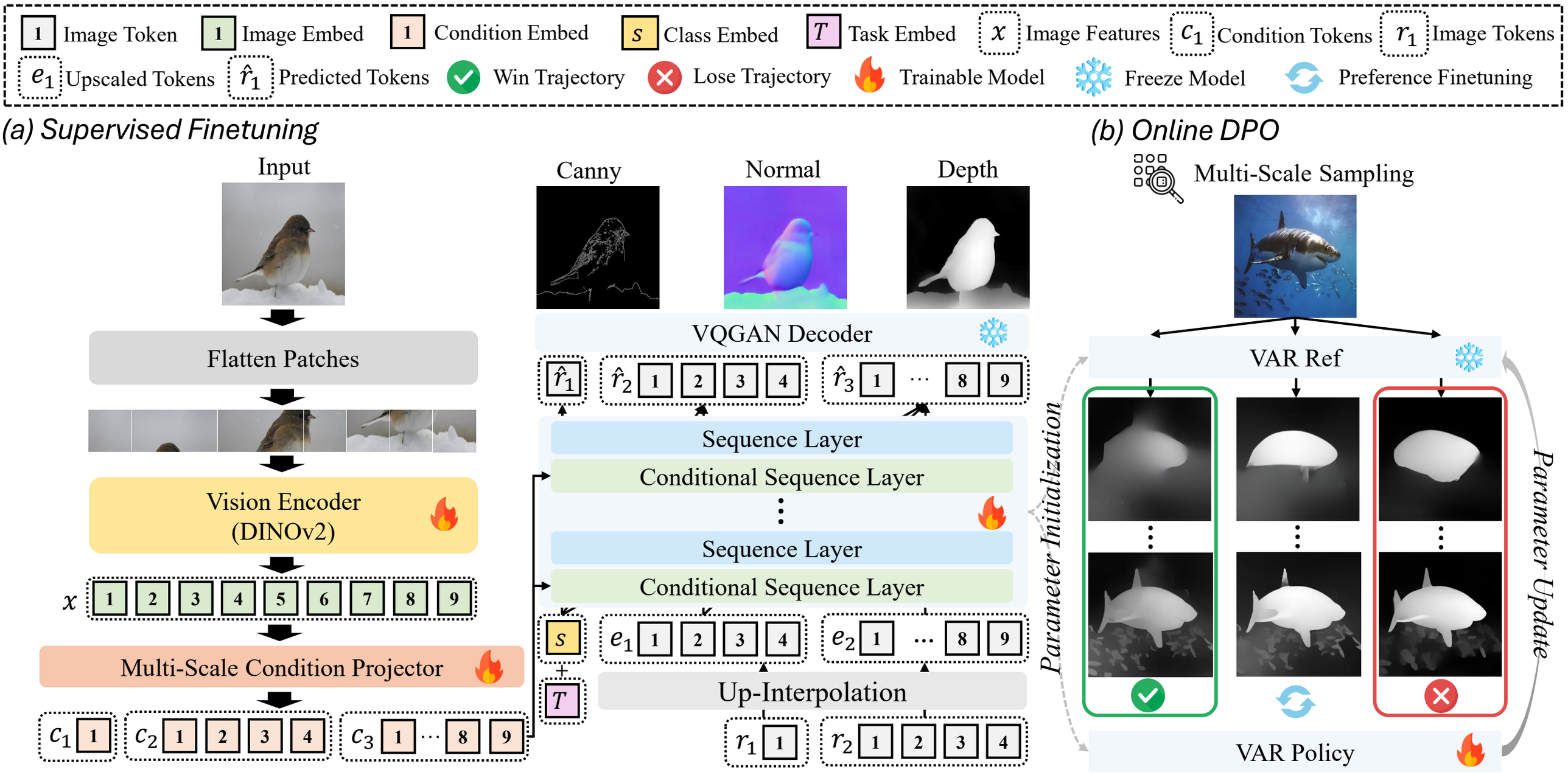
1 Image Token
1 Image Embed
1 Condition Embed
s Class Embed
T Task Embed
x Image Features
c1 Condition Tokens
r1 Image Tokens
e1 Upscaled Tokens
r̂1 Predicted Tokens
Win Trajectory
Lose Trajectory
Trainable Model
Freeze Model
Preference Finetuning
(a) Supervised Finetuning
Input
Flatten Patches
Vision Encoder (DINOv2)
x
Multi-Scale Condition Projector
c1 c2 c3
Canny
Normal
Depth
VQGAN Decoder
r̂1 r̂2 r̂3
Sequence Layer
Conditional Sequence Layer
Sequence Layer
Conditional Sequence Layer
s
T
e1 e2
Up-Interpolation
r1 r2
(b) Online DPO
Multi-Scale Sampling
VAR Ref
Parameter Initialization
Parameter Update
VAR Policy