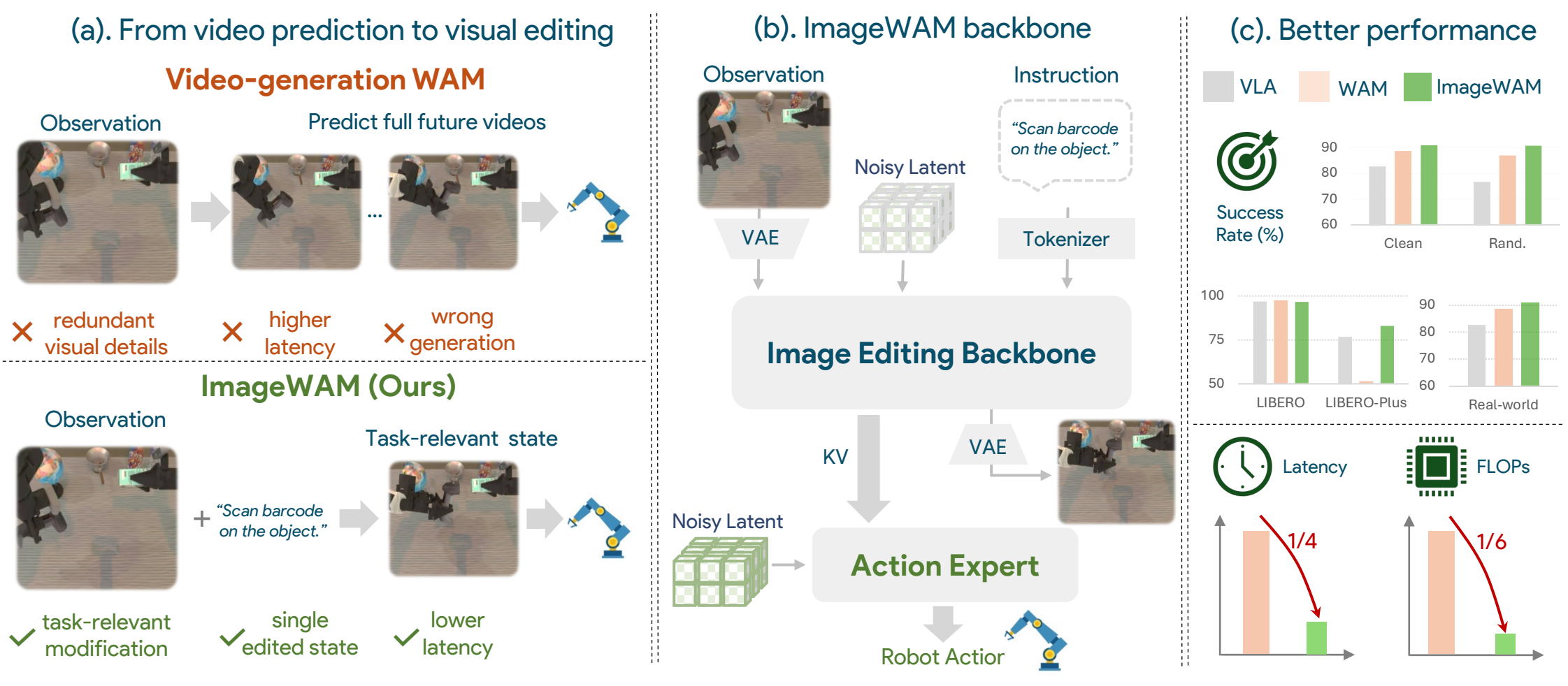
(a). From video prediction to visual editing
Video-generation WAM
Observation
Predict full future videos
redundant visual details
higher latency
wrong generation
ImageWAM (Ours)
Observation
Task-relevant state
"Scan barcode on the object."
task-relevant modification
single edited state
lower latency
(b). ImageWAM backbone
Observation
Instruction
"Scan barcode on the object."
Noisy Latent
VAE
Tokenizer
Image Editing Backbone
KV
VAE
Noisy Latent
Action Expert
Robot Actior
(c). Better performance
VLA
WAM
ImageWAM
Success Rate (%)
90
80
70
60
Clean
Rand.
100
75
50
LIBERO
LIBERO-Plus
90
80
70
60
Real-world
Latency
FLOPs
1/4
1/6