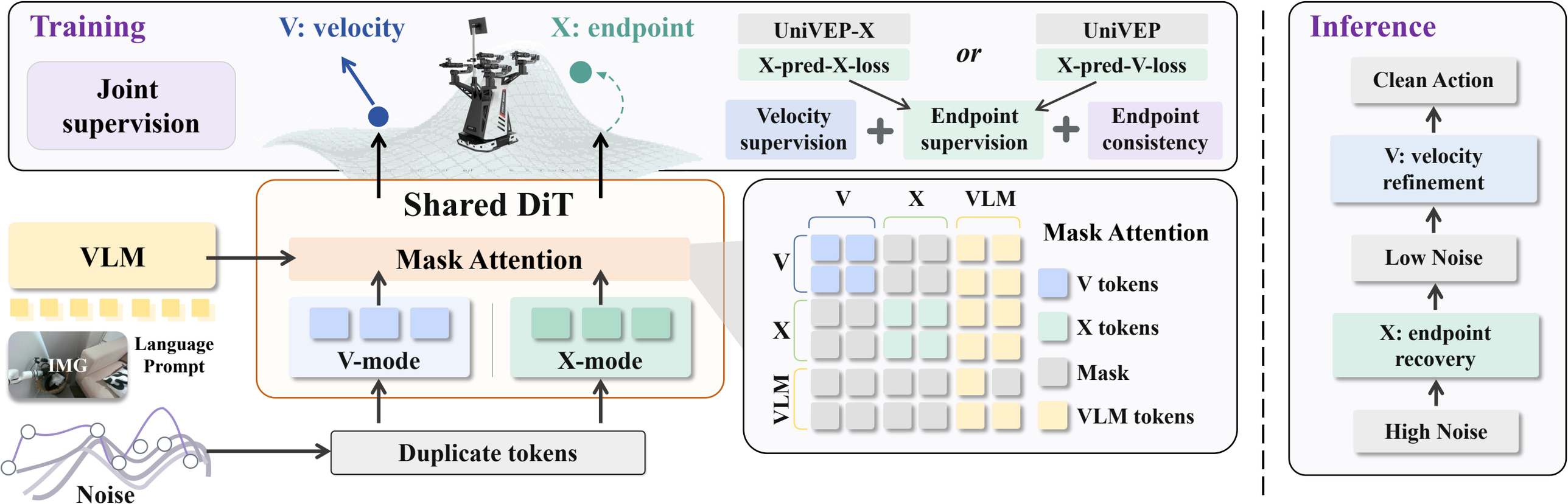

Training
V: velocity
X: endpoint
Joint supervision
UniVEP-X
X-pred-X-loss
or
UniVEP
X-pred-V-loss
Velocity supervision
Endpoint supervision
Endpoint consistency
Shared DiT
VLM
Mask Attention
V-mode
X-mode
IMG
Language Prompt
Noise
Duplicate tokens
V
X
VLM
Mask Attention
V tokens
X tokens
Mask
VLM tokens
Inference
Clean Action
V: velocity refinement
Low Noise
X: endpoint recovery
High Noise