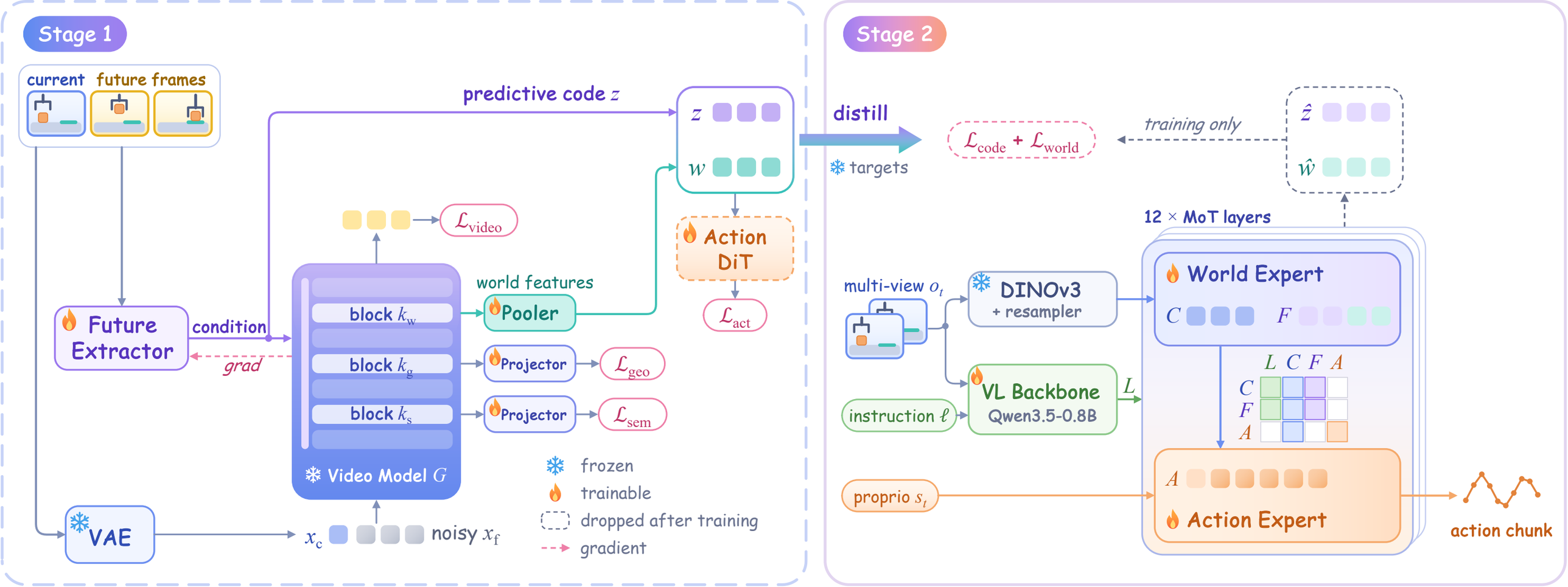
Stage 1
current
future frames
predictive code z
z
w
Action DiT
$\mathcal{L}_{\text{act}}$
$\mathcal{L}_{\text{video}}$
world features
Pooler
Future Extractor
condition
grad
block k_{w}
block k_{g}
block k_{s}
Projector
$\mathcal{L}_{\text{geo}}$
Projector
$\mathcal{L}_{\text{sem}}$
Video Model G
VAE
x_{c}
noisy x_{f}
frozen
trainable
dropped after training
gradient
Stage 2
distill
targets
$\mathcal{L}_{\text{code}} + \mathcal{L}_{\text{world}}$
training only
$\hat{z}$
$\hat{w}$
12 × MoT layers
multi-view o_t
DINOv3
+ resampler
World Expert
C
F
L C F A
C
F
A
VL Backbone
Qwen3.5-0.8B
L
instruction ℓ
proprio s_t
A
Action Expert
action chunk