Action-Object State Manipulation

Initial Frame

$\mathcal{E}$

z^0

z_t^*

©

Q-Former

Instruction: *"scooping baking soda up with spoon"*

$\hat{z}_0^*$

$\mathcal{D}$
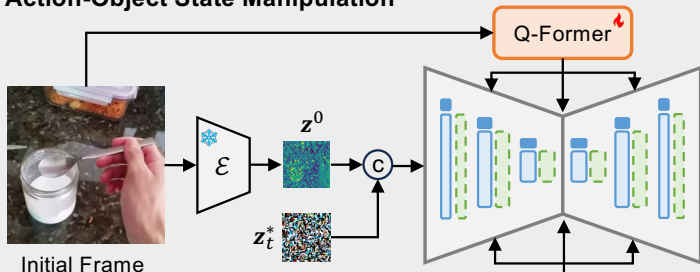
Structure Reward Tuning

Target Frame

MiDaS → $\mathcal{L}_{dp}$

HED → $\mathcal{L}_{ed}$

Depth

Edge
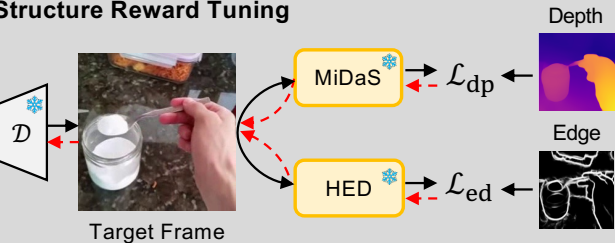
☐ Spatial Layer　■ LoRA　▢ Disabled Temporal Layer

🔥 Trainable　❄ Frozen　© Concatenate

Action-Object State Prediction

Initial Frame

$\mathcal{E}$

$[z^0; \text{MASK}]$

z_t

©

Q-Former

Instruction: *"scooping baking soda up with spoon"*

$\hat{z}_0$

Motion Reward Tuning

Q　t

P　t

$\mathcal{L}_{mo}$

Optical Flow

Downsampling

RAFT

Transition Frames

◀-- Reward Gradient
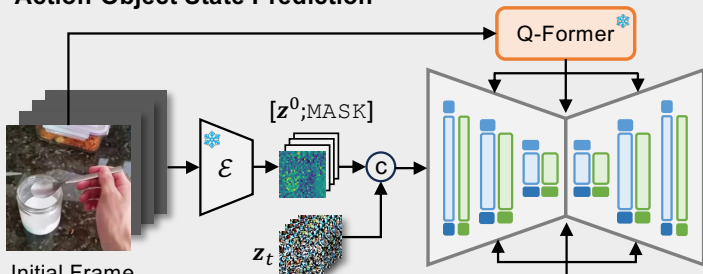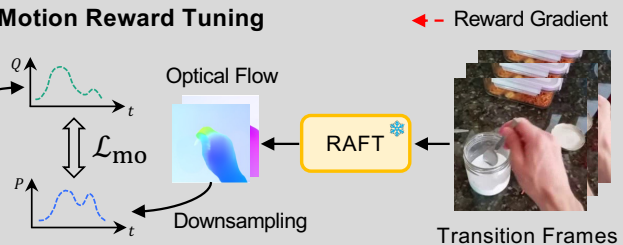
■ Spatial LoRA　▢ Activated Temporal Layer　■ Temporal LoRA