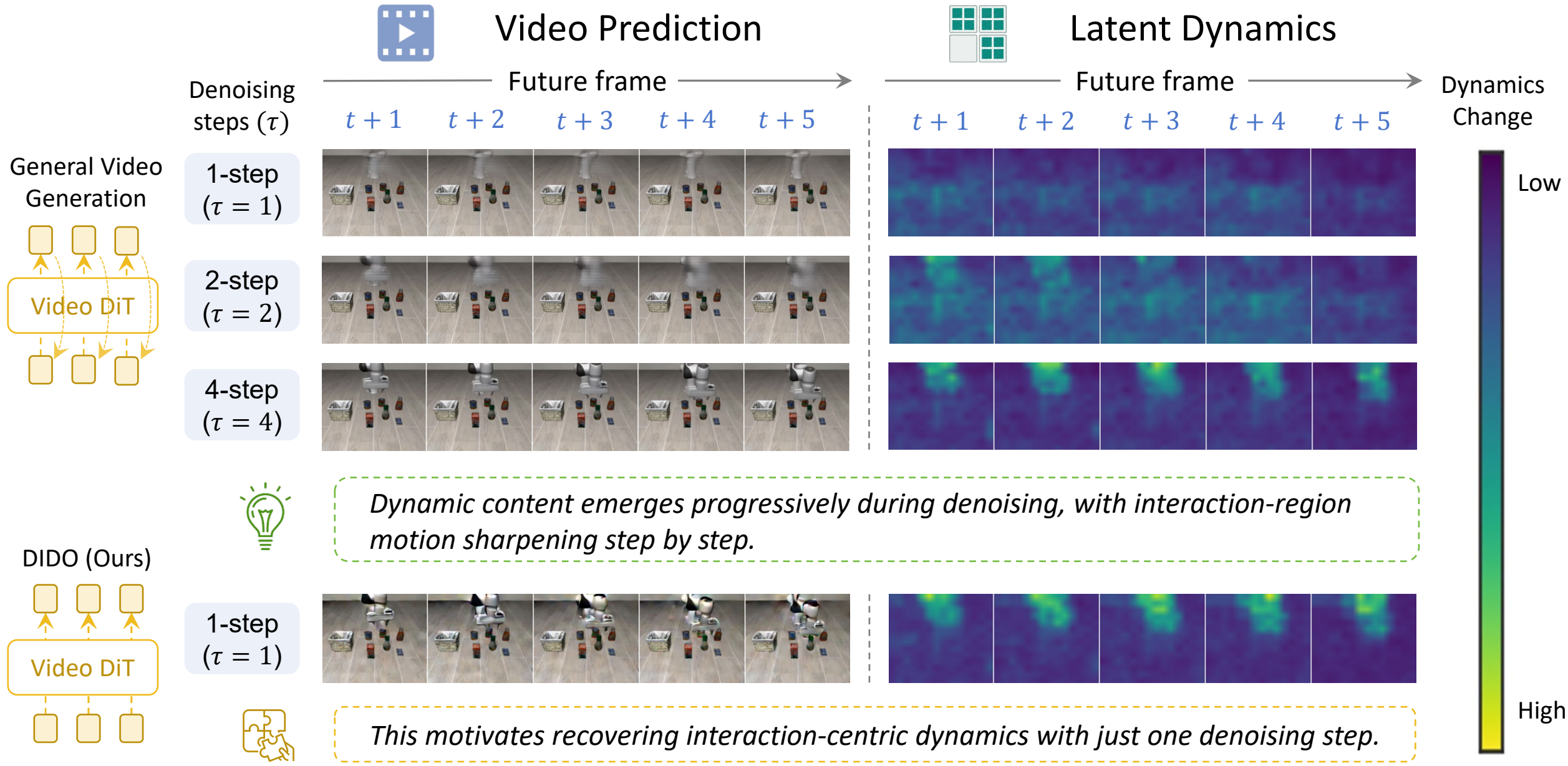
Video Prediction
Latent Dynamics
Denoising steps (τ)
Future frame
Future frame
$t+1$ $t+2$ $t+3$ $t+4$ $t+5$
$t+1$ $t+2$ $t+3$ $t+4$ $t+5$
Dynamics Change
Low
High
General Video Generation
Video DiT
1-step ($\tau = 1$)
2-step ($\tau = 2$)
4-step ($\tau = 4$)
Dynamic content emerges progressively during denoising, with interaction-region motion sharpening step by step.
DIDO (Ours)
Video DiT
1-step ($\tau = 1$)
This motivates recovering interaction-centric dynamics with just one denoising step.