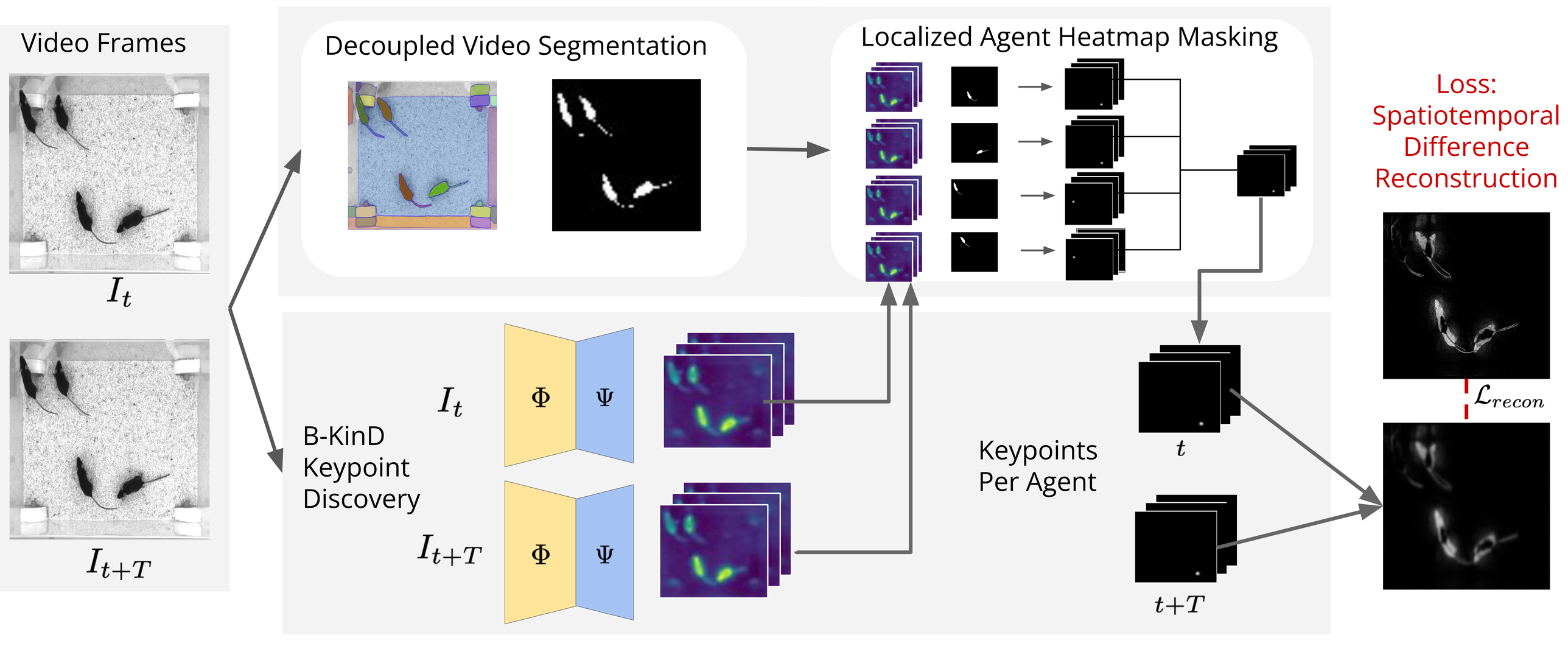
Video Frames
I_t
I_{t+T}
Decoupled Video Segmentation
Localized Agent Heatmap Masking
B-KinD
Keypoint
Discovery
I_t
I_{t+T}
Φ
Ψ
Φ
Ψ
Keypoints
Per Agent
t
$t+T$
Loss:
Spatiotemporal
Difference
Reconstruction
$\mathcal{L}_{recon}$