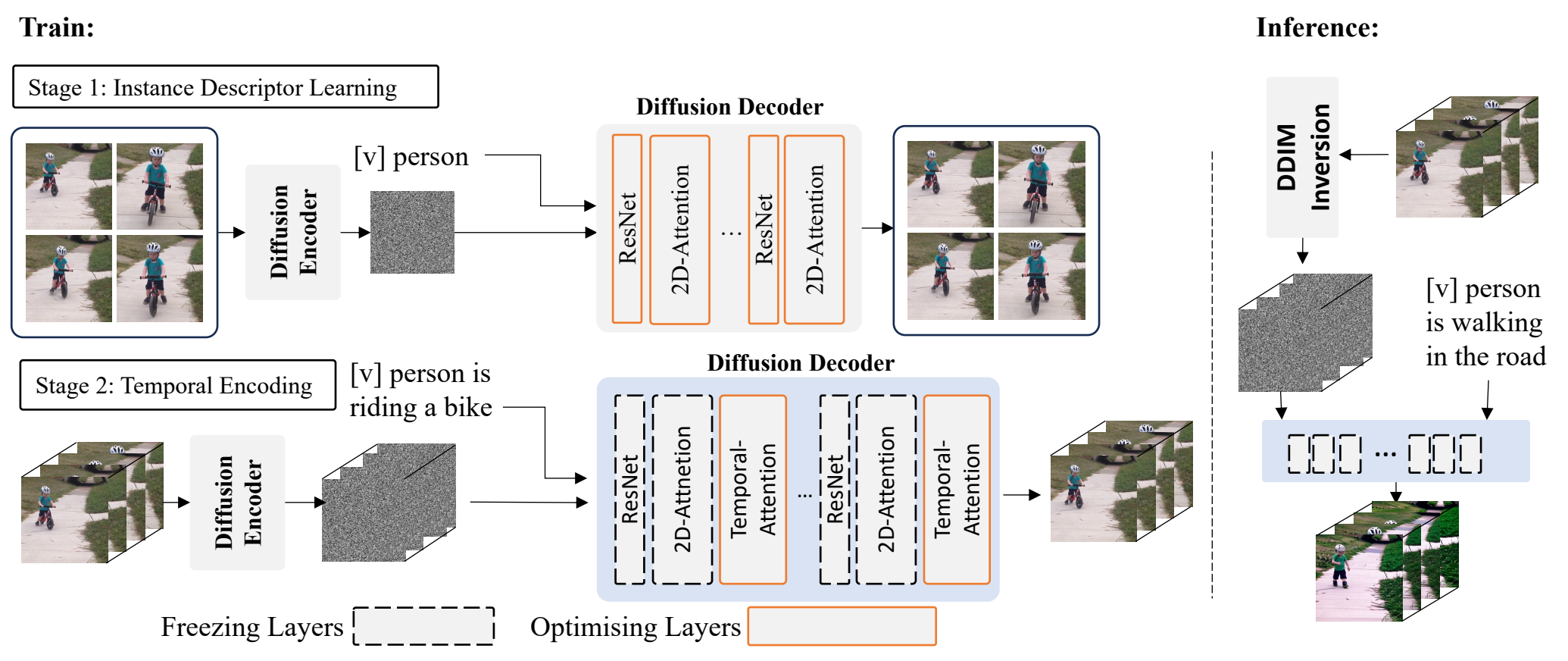
Train:
Inference:
Stage 1: Instance Descriptor Learning
Diffusion Encoder
[v] person
Diffusion Decoder
ResNet
2D-Attention
ResNet
2D-Attention
Stage 2: Temporal Encoding
[v] person is riding a bike
Diffusion Encoder
Diffusion Decoder
ResNet
2D-Attnetion
Temporal-Attention
ResNet
2D-Attention
Temporal-Attention
Freezing Layers
Optimising Layers
DDIM Inversion
[v] person is walking in the road