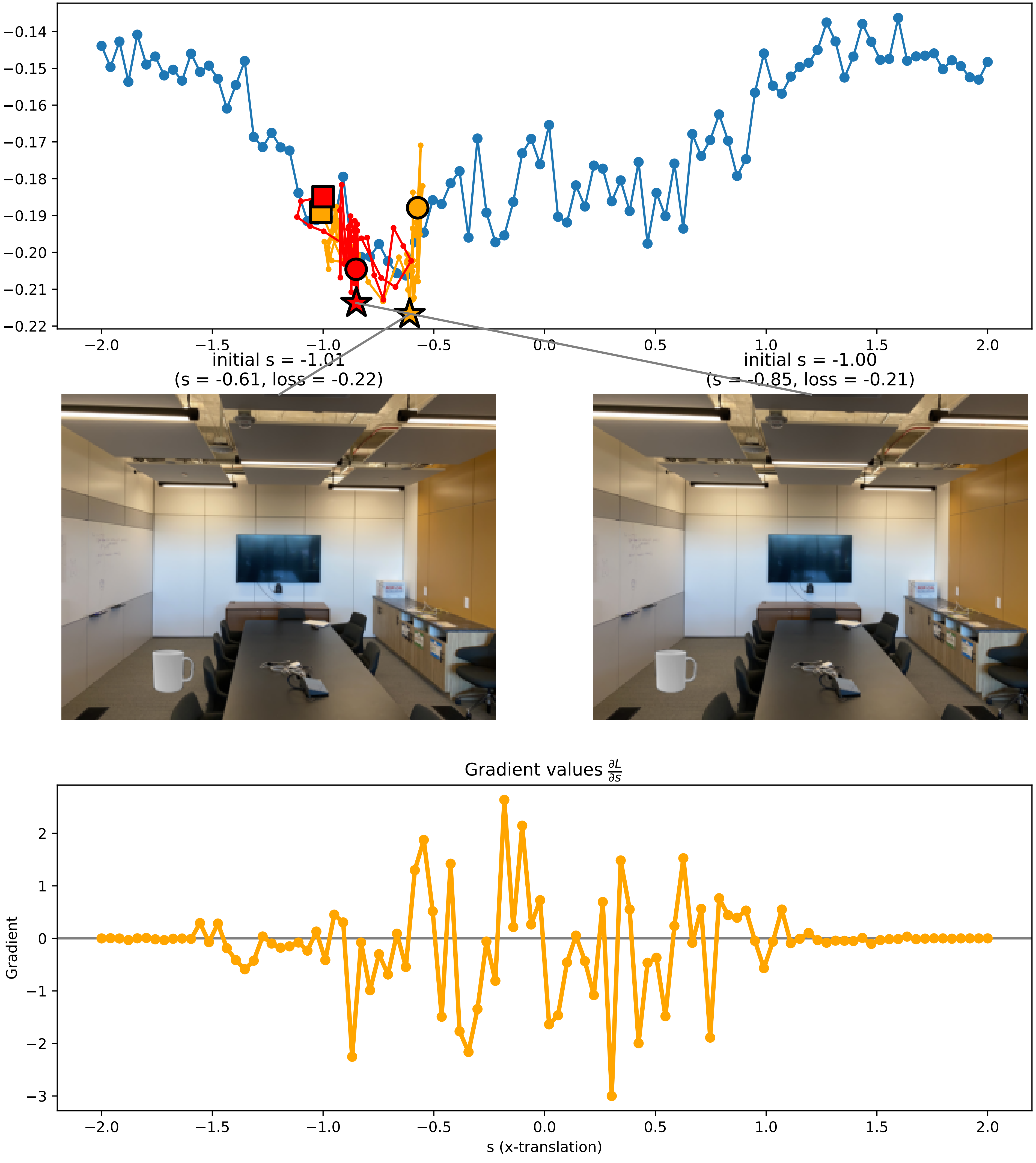
Augmented gradient descent (LR = 0.1, 8 random image augmentations, grad clip = 3.0)
prompts = ['a coffee mug on a table in an office', 'a cup of coffee on a table in an office', 'a white mug on a table in an office']
(CLIP version = ViT-L/14)


initial s = -1.01
(s = -0.61, loss = -0.22)
initial s = -1.00
(s = -0.85, loss = -0.21)
Gradient values $\frac{\partial L}{\partial s}$
Gradient
s (x-translation)