

Concat
RGBD

KL-E

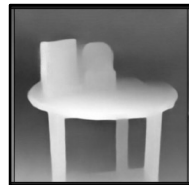
"A table with a book"

Frozen
text E

Diffusion
U-Net

KL-D

RGBD



Inference