

"a corgi  
playing a  
flame  
throwing  
trumpet"

text  
encoder

CLIP objective

img  
encoder

prior

decoder

