

Encode differences in embeddings, such as size!



"A clock that is on the side of a building"



"there are many different clocks on this wall"

CLIP Image Encoder

Difference in image embeddings

Align with  
Contrastive Loss

Embedding of text difference

Prompt a LLM:

"Q: What is the visual difference between an image of {} and an image of {}?"

Text Description of  
Difference between Images

"A: The clock on the building is much larger than the clocks on the wall"

CLIP Text Encoder