

"two cats resting in the cushion"



Retrieval



Association:  
Image Retrieval

"a black kitten"



Grounding



Instance-level Association:  
Textual Grounding on Images/Videos  
(also known as referring expression comprehension)



Captioning  
& QA

"two cats resting in the cushion"

"How many cats ...?"  
-> "two"

Interpretation & Understanding:  
Image Caption, VQA