text concept extraction
cross-modality consistency
vision concept extraction
"A dog and a cow on grass."
concept identification
"A dog and a cow on grass."
text encoder
vision encoder
linear classifier
text-guided pooling
"dog"
"cow"
"grass"
text representation
vision representation
a.
b.
c.