supplementary

October 17, 2024

1 Examples of failed cases:

TThe example shows that while both models retrieved correct images—i.e., images of streets—one of the images retrieved by SR_{clip} was marked incorrect during evaluation because it lacked "street" in its labels. Since we rely on the provided labels for evaluating model performance, this led to a mismatch. There are several similar cases in the Conceptual Captions dataset, where images contain specific objects or features that are missing from the labels. This is likely due to the large number of labels in the dataset, inaccuracies in the word mapping, and the fact that the labels are not human-generated.

['neighbourhood', 'town', 'urban area', 'city', 'downtown', 'street', 'metropolitan area', 'human settlement', 'residential area', 'road', 'lane', 'building', 'mixed-use', 'transport', 'mode of transport']
['residential area', 'neighbourhood', 'town', 'urban area', 'city', 'metropolitan area', 'human settlement', 'lane', 'street', 'road', 'building', 'mixed-use', 'architecture', 'suburb', 'downtown']
['residential area', 'town', 'neighbourhood', 'building', 'property', 'human settlement', 'street', 'mixed-use', 'city', 'downtown', 'road', 'architecture', 'urban area', 'house', 'apartment']

['street', 'town', 'city', 'building', 'neighbourhood', 'architecture', 'urban area', 'landmark', 'human settlement', 'sky', 'road', 'metropolis', 'infrastructure', 'metropolitan area', 'downtown']
['sky', 'urban area', 'city', 'metropolitan area', 'daytime', 'vehicle', 'human settlement', 'mode of transport', 'town', 'cloud', 'road', 'transport', 'architecture', 'car', 'metropolis']
['town', 'neighbourhood', 'vehicle', 'transport', 'street', 'mode of transport', 'car', 'human settlement', 'city' 'thoroughfare', 'urban area', 'downtown', 'road', 'lane', 'pedestrian']

Figure 1: **Retrieval images and their labels for Conceptual Captions dataset for Query "street without trees" using CLIP and SR_{clip}**