Examples of failed cases:

Exclusion based query

The 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.

['**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']

['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']

Figure 1: **For the query "street without trees" in the Conceptual Captions dataset, the following images and their labels were retrieved. The first set shows the three closest images using SR_{blip} embeddings, along with their corresponding labels. The second set shows the nearest images retrieved using BLIP embeddings. All the images retrieved by SR_{blip} contain a street, but the label for the second image does not include the word "street." In contrast, the label "street" appears in all the images retrieved by BLIP embeddings which makes the images correct according to the labels.**

Normal query

We observe a similar issue when retrieving images for the query "landscape." While the retrieved images from both representations do contain landscapes, some of the labels for images retrieved by SR_{blip} include terms like "natural

landscape" and don't include "landscape". This inconsistency in labels is why even correctly retrieved images may result in poor evaluations when compared to the ground truth labels.

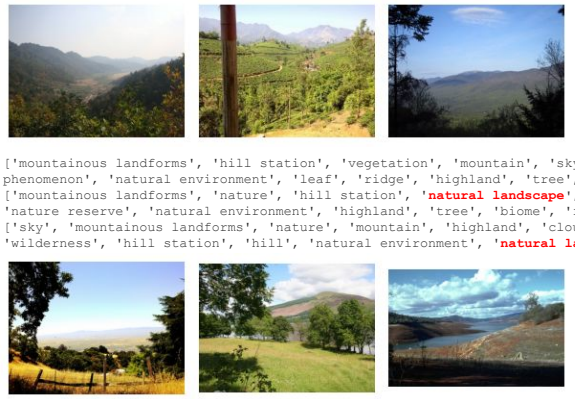

Figure 2: **For the query "landscape" in the Conceptual Captions dataset, images and their labels were retrieved. The first set presents the three nearest images using SR_{blip} embeddings, along with their labels. The second set shows the nearest images retrieved using BLIP embeddings. Although all the images retrieved by SR_{blip} contain landscapes, the labels for the second and third images include "natural landscape", which makes these retrieved images incorrect according to the ground truth**