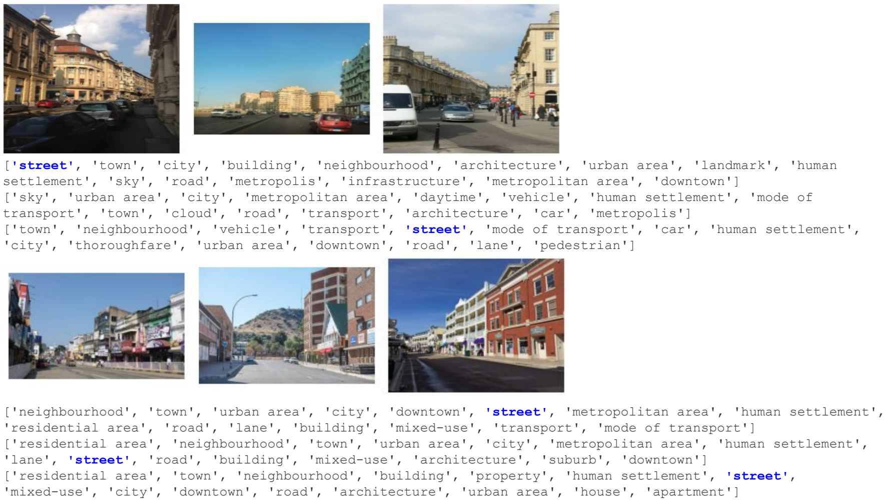
Examples of failed cases for Conceptual captions Dataset:

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.



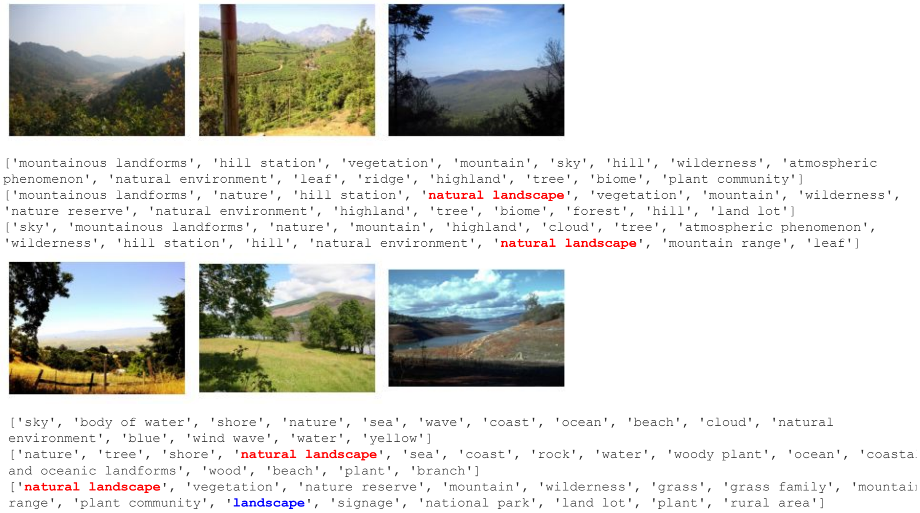
```
['street', 'town', 'city', 'building', 'neighbourhood', 'architecture', 'urban area', 'landmark', 'human
settlement', 'sky', 'road', 'metropolis', 'infrastructure', 'metropolitan area', 'downtown']
['sky', 'urban area', 'city', 'metropolitan area', 'daytime', 'vehicle', 'human settlement', 'mode 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.



```
['mountainous landforms', 'hill station', 'vegetation', 'mountain', 'sky', 'hill', 'wilderness', 'atmospheric
phenomenon', 'natural environment', 'leaf', 'ridge', 'highland', 'tree', 'biome', 'plant community']
['mountainous landforms', 'nature', 'hill station', 'natural landscape', 'vegetation', 'mountain', 'wilderness',
'nature reserve', 'natural environment', 'highland', 'tree', 'biome', 'forest', 'hill', 'land lot']
['sky', 'mountainous landforms', 'nature', 'mountain', 'highland', 'cloud', 'tree', 'atmospheric phenomenon',
'wilderness', 'hill station', 'hill', 'natural environment', 'natural landscape', 'mountain range', 'leaf']
```

```
['sky', 'body of water', 'shore', 'nature', 'sea', 'wave', 'coast', 'ocean', 'beach', 'cloud', 'natural
environment', 'blue', 'wind wave', 'water', 'yellow']
['nature', 'tree', 'shore', 'natural landscape', 'sea', 'coast', 'rock', 'water', 'woody plant', 'ocean', 'coasta
and oceanic landforms', 'wood', 'beach', 'plant', 'branch']
['natural landscape', 'vegetation', 'nature reserve', 'mountain', 'wilderness', 'grass', 'grass family', 'mountai
range', 'plant community', 'landscape', 'signage', 'national park', 'land lot', 'plant', 'rural area']
```

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**

0.1 Qualitative Evaluation

0.1.1 Exclusion Based Retrieval:

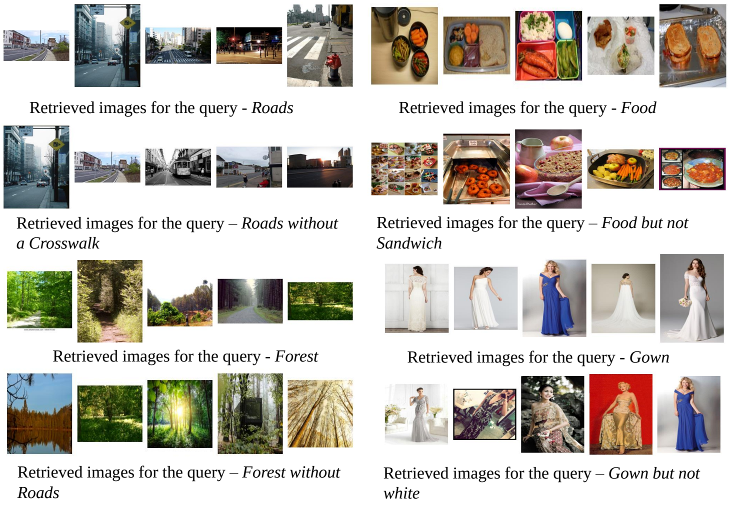
We next conduct a qualitative evaluation of our representations to examine their effectiveness in exclusion based retrieval scenerio through several examples. Fig-3 presents two examples from the MSCOCO and Conceptual Captions datasets to demonstrate exclusion-based retrieval. In the first image of the figure, the first row displays the top 5 images retrieved for the query *Roads*, where the 3rd and 5th images include crosswalks. The objective in exclusion-based retrieval is to return images of roads without crosswalks. The second row shows the results using SR_{clip} representations, which successfully retrieve images of roads that do not include crosswalks.

0.1.2 Disentanglement:

Our method enables effective disentanglement of image and text representations at the dimension level, where similar concepts activate similar dimensions with high values. The model produces sparse embeddings, ensuring that specific dimensions correspond to meaningful features. In Figure 5, we illustrate this by presenting a query, a ranked list of top k% of active dimensions in both the query embedding with common dimensions being shown in bold and the top five retrieved images using only the dimensions in the list. We further demonstrate that query pairs in the same row exhibit similarities, as they share common active dimensions. This supports the idea that the activated dimensions capture essential features and semantic factors. For instance, for the query ”valley”, the embeddings of the top five retrieved images consistently highlight specific high-value dimensions, as shown in the corresponding list in the figure. Similarly, the query ”trees” activates a distinct set of dimensions, yet some overlap exists between ”valley” and ”trees”, indicating shared semantic attributes.

Additionally, our model effectively captures semantically related concepts within datasets. To showcase this, we present examples from the Conceptual Captions dataset in Figure 4. This dataset pairs web images with their associated captions, often omitting crucial details that are present in the images. As a result, when retrieving images using CLIP embeddings, the returned images and captions frequently fail to align with the actual content. However, this issue is significantly mitigated with our SR_{clip} embeddings, which enhance retrieval accuracy by focusing on contextually relevant features. In Figure 4, we compare retrieval results for the queries ”Animals” and ”Food”, displaying the most similar images alongside the most frequent words in the top retrieved sentences. We also contrast these results with retrievals using CLIP embeddings for the queries ”images of animals” and ”images of food”. Notably, the words retrieved using SR_{Clip} are more relevant and closely aligned with the query. This highlights the disentanglement capability of our embeddings, as they effectively separate and organize data based on similar concepts and words, leading to more accurate and contextually appropriate retrieval results in such datasets.

Exclusion Retrieval Qualitative Evaluation Examples



Retrieved images for the query - *Roads*

Retrieved images for the query - *Food*

Retrieved images for the query – *Roads without a Crosswalk*

Retrieved images for the query – *Food but not Sandwich*

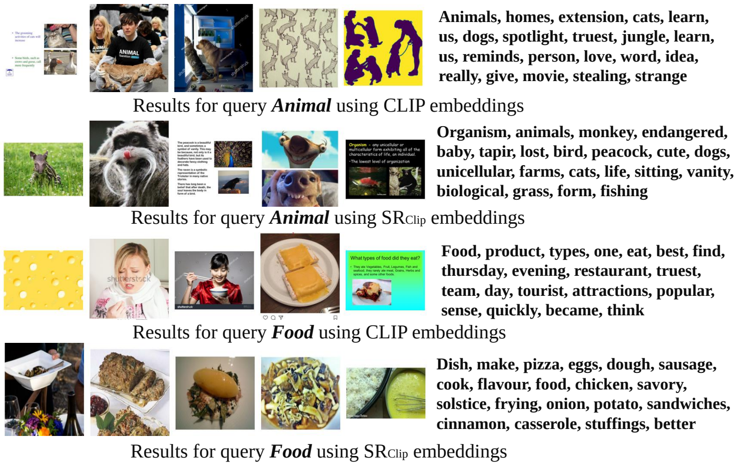
Retrieved images for the query - *Forest*

Retrieved images for the query - *Gown*

Retrieved images for the query – *Forest without Roads*

Retrieved images for the query – *Gown but not white*

Figure 3: Exclusion based Retrieval examples from MSCOCO and Conceptual Captions datasets using SR_{clip}



Animals, homes, extension, cats, learn, us, dogs, spotlight, truest, jungle, learn, us, reminds, person, love, word, idea, really, give, movie, stealing, strange

Results for query ***Animal*** using CLIP embeddings

Organism, animals, monkey, endangered, baby, tapir, lost, bird, peacock, cute, dogs, unicellular, farms, cats, life, sitting, vanity, biological, grass, form, fishing

Results for query ***Animal*** using SR_{Clip} embeddings

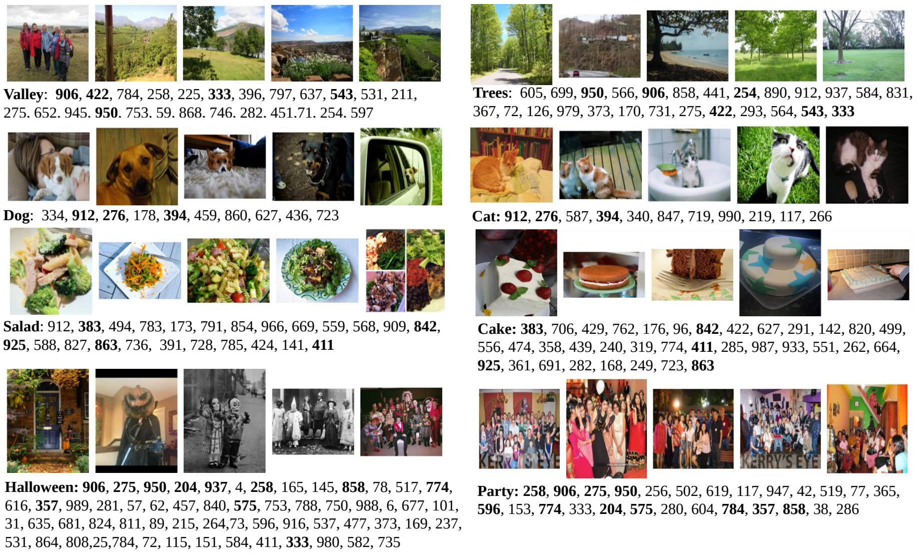
Food, product, types, one, eat, best, find, thursday, evening, restaurant, truest, team, day, tourist, attractions, popular, sense, quickly, became, think

Results for query ***Food*** using CLIP embeddings

Dish, make, pizza, eggs, dough, sausage, cook, flavour, food, chicken, savory, solstice, frying, onion, potato, sandwiches, cinnamon, casserole, stuffings, better

Results for query ***Food*** using SR_{Clip} embeddings

Figure 4: Top Retrieved Images and most frequent words from the top retrieved texts from Conceptual Captions dataset



Valley: **906**, **422**, 784, 258, 225, **333**, 396, 797, 637, **543**, 531, 211, 275. 652. 945. **950**. 753. 59. 868. 746. 282. 451.71. 254. 597

Trees: 605, 699, **950**, 566, **906**, 858, 441, **254**, 890, 912, 937, 584, 831, 367, 72, 126, 979, 373, 170, 731, 275, **422**, 293, 564, **543**, **333**

Dog: 334, **912**, **276**, 178, **394**, 459, 860, 627, 436, 723

Cat: 912, **276**, 587, **394**, 340, 847, 719, 990, 219, 117, 266

Salad: 912, **383**, 494, 783, 173, 791, 854, 966, 669, 559, 568, 909, **842**, **925**, 588, 827, **863**, 736, 391, 728, 785, 424, 141, **411**

Cake: 383, 706, 429, 762, 176, 96, **842**, 422, 627, 291, 142, 820, 499, 556, 474, 358, 439, 240, 319, 774, **411**, 285, 987, 933, 551, 262, 664, **925**, 361, 691, 282, 168, 249, 723, **863**

Halloween: 906, **275**, **950**, **204**, **937**, 4, **258**, 165, 145, **858**, 78, 517, **774**, 616, **357**, 989, 281, 57, 62, 457, 840, **575**, 753, 788, 750, 988, 6, 677, 101, 31, 635, 681, 824, 811, 89, 215, 264,73, 596, 916, 537, 477, 373, 169, 237, 531, 864, 808,25,784, 72, 115, 151, 584, 411, **333**, 980, 582, 735

Party: 258, **906**, **275**, **950**, 256, 502, 619, 117, 947, 42, 519, 77, 365, **596**, 153, **774**, 333, **204**, **575**, 280, 604, **784**, **357**, **858**, 38, 286

Figure 5: Top Retrieved Images and list of dimensions having highest values in decreasing order for given query words