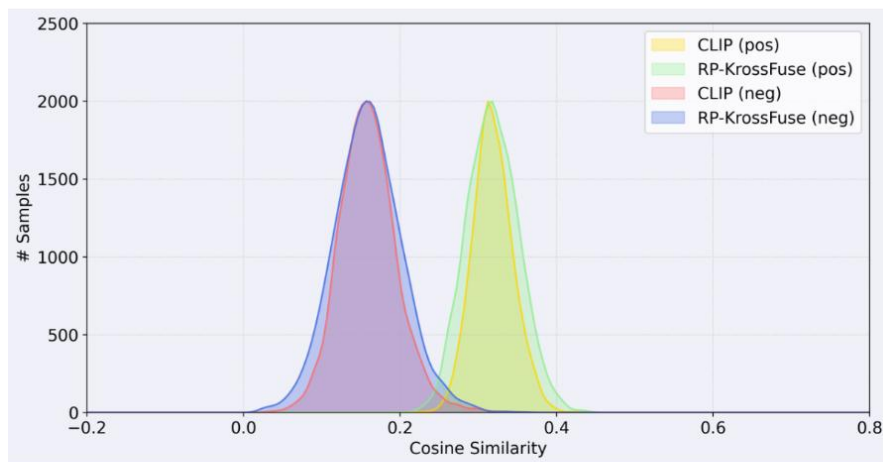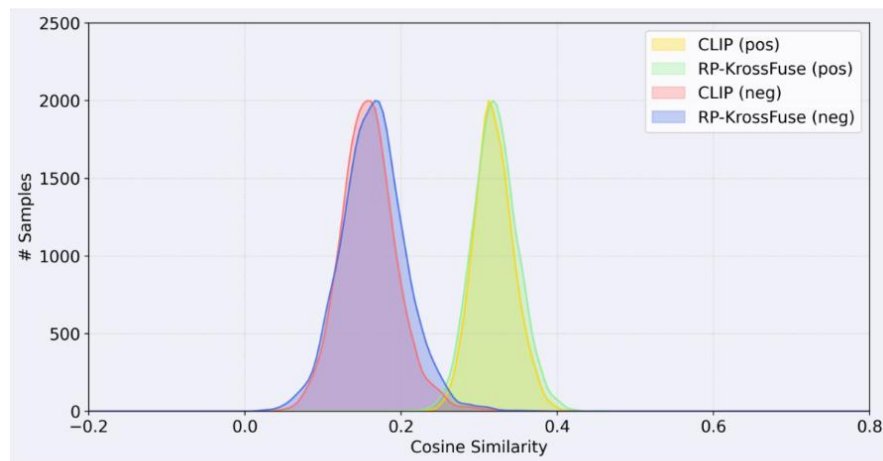C=0.1

C=0.5

C=1.0

Fig1. Pairwise cosine similarities of positive/negative samples in COCO with different values of hyperparameter C. (Response to Review 29Mf, Item 7.)

Joint Cluster 1 (mismatched image and text)

01. There is a phone on top of a calculator
02. A infant holding a baby toothbrush in his hand looking at it.
03. A little boy is playing with a suitcase.
04. A young boy reaching up to grab a red apple.

Joint Cluster 2 (image only, missing text)

Joint Cluster 3 (text only, missing image)

01. A herd of sheep grazing on a lush green field.

02. A dog driving an SUV in an open grass covered field.

03. A man that is standing in the grass near animals.

04. A cow grazing on the grass on the side of a side walk.

05. A lush green forest filled with lots of trees.

06. A horse standing in a lush green grass field.

07. A mother and it's calf standing in a field.

08. a number of horses in a field near a building

09. Two horses grazing on grass in the sun

Joint Cluster 4 (text only, missing image)

01. there is a cat that is drinking out of the toilet

02. A cat sitting in a bathroom sink under a faucet.

03. A small cat sits perched on the back of a toilet.

04. A black cat is sitting in a bathroom sink.

05. A black cat sits in a bathroom sink.

06. a cat that is standing on a toilet and next to a sink

07. A calico cat drinking from a sink faucet.

08. a cat sitting in a bathtub not looking amused

09. An orange and white cat laying on top of a sink.

Fig2. The failure case of Naive Concatenation in joint clustering results on the COCO dataset. (Response to Review M1uF, Item 1.)