# Human in the Loop

Example-1



caption\_0: the happy one is on the left and the upset one is on the right

caption\_1: the upset one is on the left and the happy one is on the right

Human: I can easily pick right combination. I looked at subject and object (direction) and I got right answer

CLIP image-text match scores:

image\_0, caption\_0: 25.416194915771484

image\_0, caption\_1: 25.222095489501953

image\_1, caption\_0: 21.922069549560547

image\_1, caption\_1: 20.996662139892578

Caption prediction failed

Image prediction failed

Group prediction failed

Computer:

Example-2

A collage of a person and person

Description automatically generated with low confidence

caption\_0: fake mustache and real hair on head

caption\_1: real mustache and fake hair on head

Human: I got it right as I understood person’s hair is wig, but this can be crazy hard for computer as wig sits exactly like real hair

CLIP image-text match scores:

image\_0, caption\_0: 22.83394432067871

image\_0, caption\_1: 23.49127197265625

image\_1, caption\_0: 24.516468048095703

image\_1, caption\_1: 24.260835647583008

Caption prediction failed

Image prediction failed

Group prediction failed

Computer:

Example-3

A picture containing text, mammal, domestic cat

Description automatically generated

caption\_0: real cat with toy baby

caption\_1: toy cat with real baby

Human: I got it right. In this case bounding boxes can help (real-cat, baby-toy)

CLIP image-text match scores:

image\_0, caption\_0: 30.87631607055664

image\_0, caption\_1: 30.757173538208008

image\_1, caption\_0: 24.18267059326172

image\_1, caption\_1: 25.401016235351562

Caption is correctly predicted

Image prediction failed

Group prediction failed

Example-4

A group of lions

Description automatically generated with low confidence

caption\_0: the lion is loud and the lionness is quiet

caption\_1: the lion is quiet and the lionness is loud

Human: I got it right, I think bounding boxes will help (find lion then its emotion). In 2 predicate cases associating subject with object will also help

CLIP image-text match scores:

image\_0, caption\_0: 28.72422218322754

image\_0, caption\_1: 28.407772064208984

image\_1, caption\_0: 27.071542739868164

image\_1, caption\_1: 26.811992645263672

Caption prediction failed

Image prediction failed

Group prediction failed

Eample-5

A picture containing text, room

Description automatically generated

caption\_0: there is a pink couch with green pillows

caption\_1: there is a green couch with pink pillows

Human, I got it right by differentiating pillow from couch. I think association is failing here. Similar cases are also above

CLIP image-text match scores:

image\_0, caption\_0: 28.59800148010254

image\_0, caption\_1: 28.69970703125

image\_1, caption\_0: 28.189571380615234

image\_1, caption\_1: 28.39978790283203

Caption prediction failed

Image prediction failed

Group prediction failed

Example-6

A dog and a cat sitting on a couch

Description automatically generated with medium confidence

caption\_0: light colored dog with dark colored cat

caption\_1: dark colored dog with light colored cat

Human:I got it right. Similar failure of association (light colour – dog (caption-1))

CLIP image-text match scores:

image\_0, caption\_0: 26.476428985595703

image\_0, caption\_1: 25.83609962463379

image\_1, caption\_0: 30.660165786743164

image\_1, caption\_1: 30.456029891967773

Caption prediction failed

Image prediction failed

Group prediction failed

Example-7

A person walking a dog

Description automatically generated with medium confidence

caption\_0: someone is wearing a black hat and yellow shirt

caption\_1: someone is wearing a yellow hat and black shirt

CLIP image-text match scores:

image\_0, caption\_0: 25.040931701660156

image\_0, caption\_1: 25.583885192871094

image\_1, caption\_0: 19.73490333557129

image\_1, caption\_1: 19.47063636779785

Caption prediction failed

Image prediction failed

Group prediction failed

Human: got it right. Bounding boxes around mentioned objects would help

Example-8:

A picture containing text

Description automatically generated

caption\_0: black socks and orange shoes

caption\_1: orange socks and black shoes

Human: I got it right. Associating words will help.

CLIP image-text match scores:

image\_0, caption\_0: 26.201927185058594

image\_0, caption\_1: 25.871469497680664

image\_1, caption\_0: 28.376188278198242

image\_1, caption\_1: 28.013500213623047

Caption prediction failed

Image prediction failed

Group prediction failed

Example -9

A picture containing graphical user interface

Description automatically generated

caption\_0: red building with white shutters

caption\_1: white building with red shutters

Human: I got it right but computer can confuse as lot of unrelated objects in second image.

CLIP image-text match scores:

image\_0, caption\_0: 28.34311294555664

image\_0, caption\_1: 29.549785614013672

image\_1, caption\_0: 23.082778930664062

image\_1, caption\_1: 23.18783950805664

Caption prediction failed

Image prediction failed

Group prediction failed

Example-10

A picture containing text, television, monitor, screen

Description automatically generated

caption\_0: blue pants and green top

caption\_1: green pants and blue top

Human:

CLIP image-text match scores:

image\_0, caption\_0: 22.58848762512207

image\_0, caption\_1: 22.848918914794922

image\_1, caption\_0: 18.017168045043945

image\_1, caption\_1: 18.105844497680664

Caption prediction failed

Image prediction failed

Group prediction failed