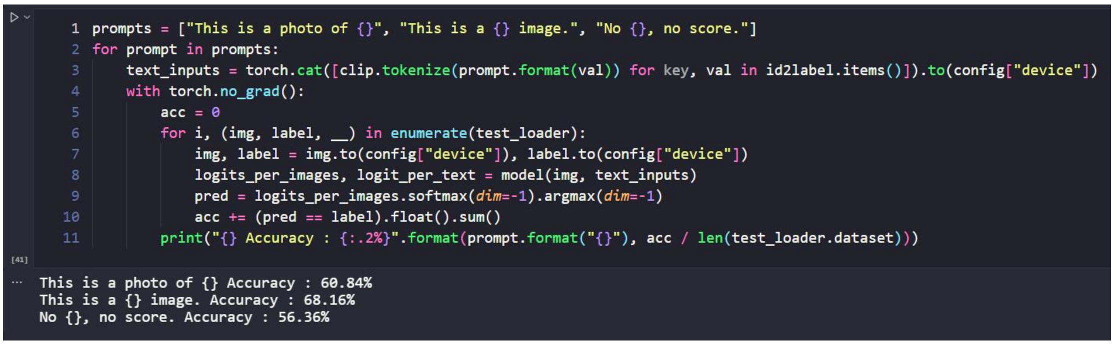
Problem 1.

1. 由於 Clip 是對比 image 與 text 間的關聯程度， 而這種配對在網路上可大量取得，且也被 Clip 作為訓練資料集，因此模型的泛化能力也較強，只要 image 與訓練集的某種分類名字有較大關連，Clip 就可以準確地分類

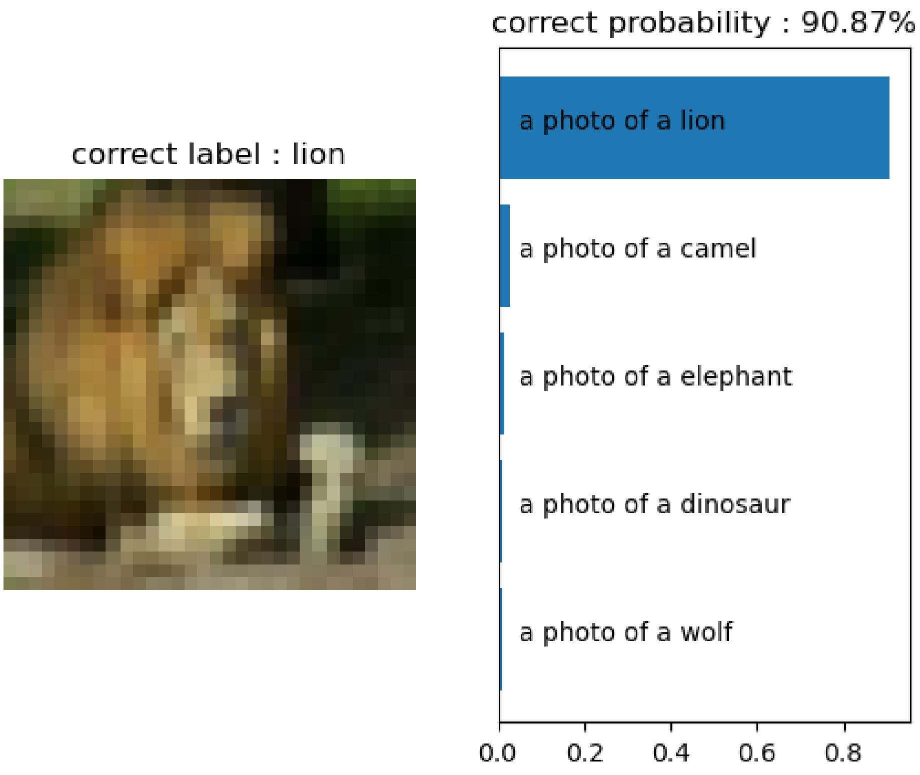
2. 使用”This is a {object} image”當作 prompt-text 時，有最高的準確率，若使用”No {object}, no score.”則最低

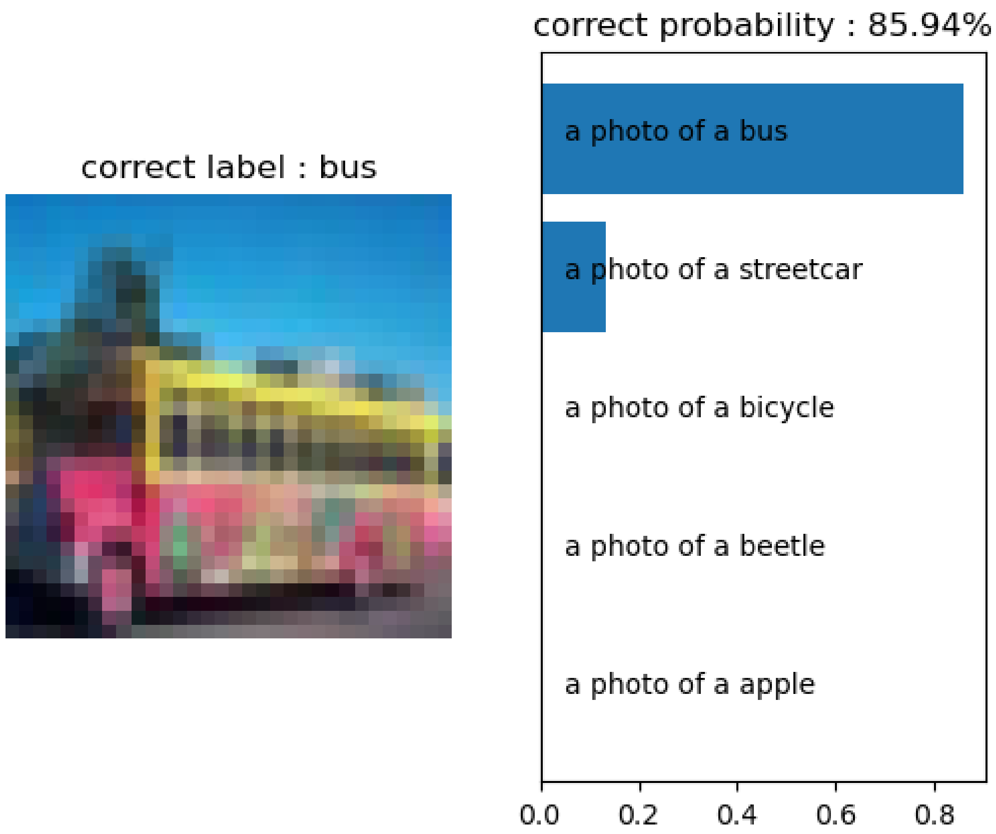


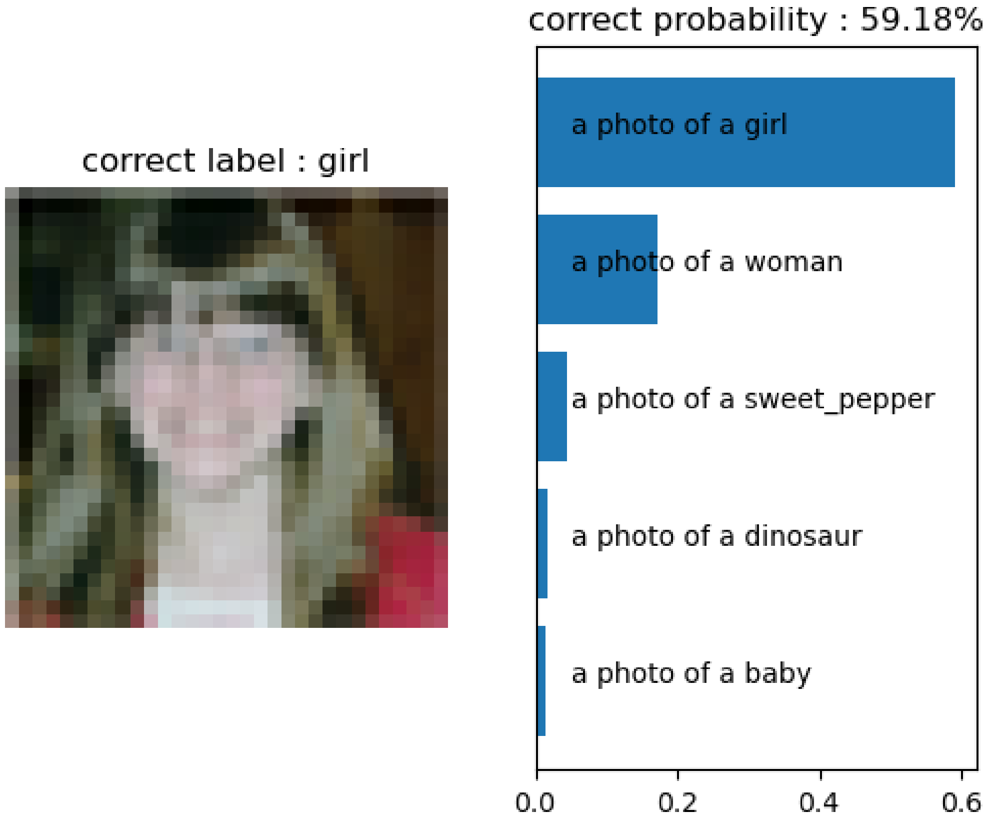
```
prompts = ["This is a photo of {}", "This is a {} image.", "No {}, no score."]
for prompt in prompts:
    text_inputs = torch.cat([clip.tokenize(prompt.format(val)) for key, val in id2label.items()]).to(config["device"])
    with torch.no_grad():
        acc = 0
        for i, (img, label, __) in enumerate(test_loader):
            img, label = img.to(config["device"]), label.to(config["device"])
            logits_per_images, logit_per_text = model(img, text_inputs)
            pred = logits_per_images.softmax(dim=-1).argmax(dim=-1)
            acc += (pred == label).float().sum()
        print("{} Accuracy : {:.2%}".format(prompt.format("{}"), acc / len(test_loader.dataset)))
```

```
This is a photo of {} Accuracy : 60.84%
This is a {} image. Accuracy : 68.16%
No {}, no score. Accuracy : 56.36%
```

3.




correct label : bus
correct probability : 85.94%
a photo of a bus
a photo of a streetcar
a photo of a bicycle
a photo of a beetle
a photo of a apple
0.0
0.2
0.4
0.6
0.8


correct label : girl
correct probability : 59.18%
a photo of a girl
a photo of a woman
a photo of a sweet_pepper
a photo of a dinosaur
a photo of a baby
0.0
0.2
0.4
0.6

Problem 2.

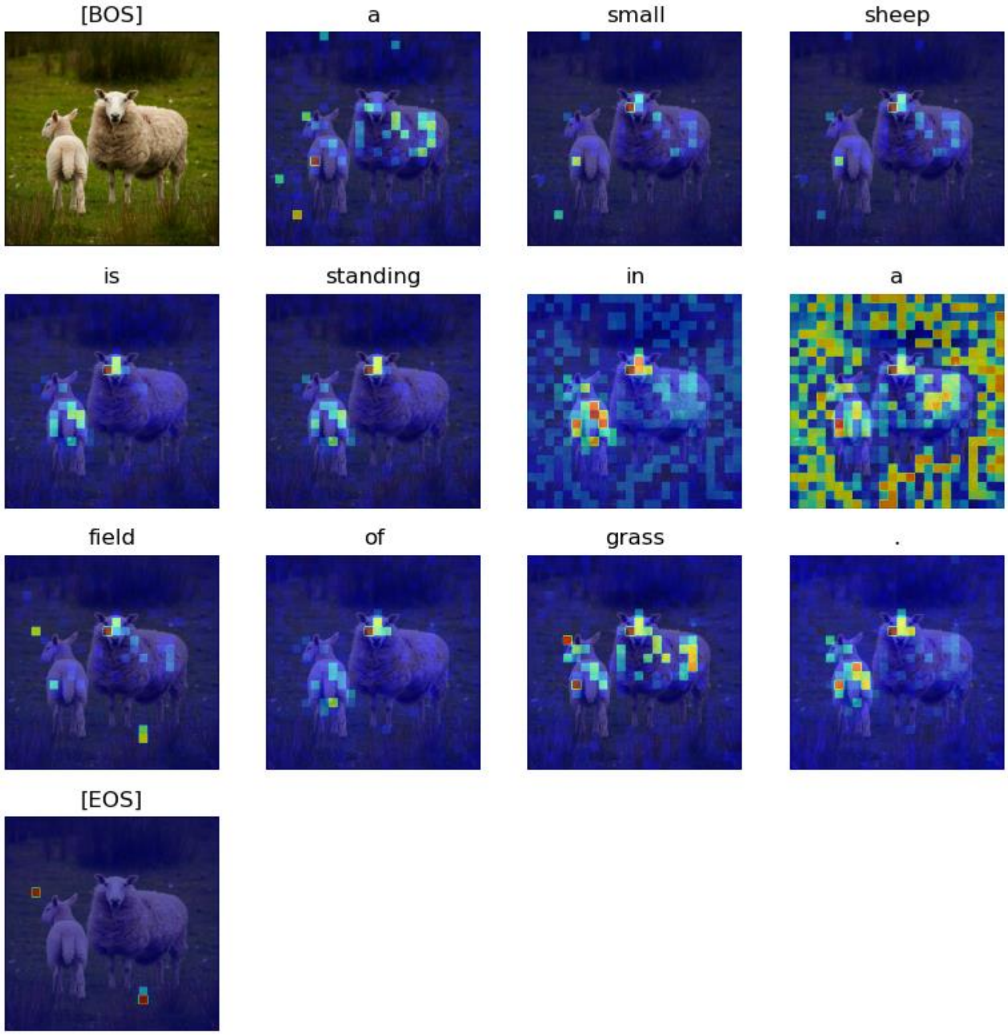
1. CIDEr : 0.892, CLIPScore : 0.710

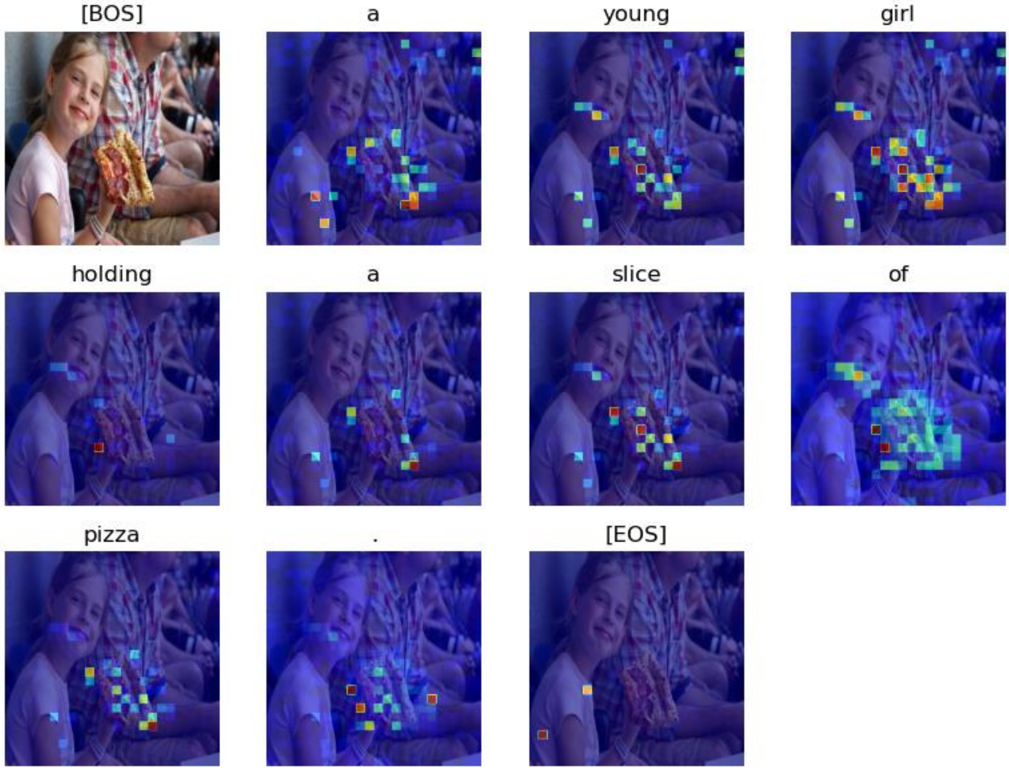
2.

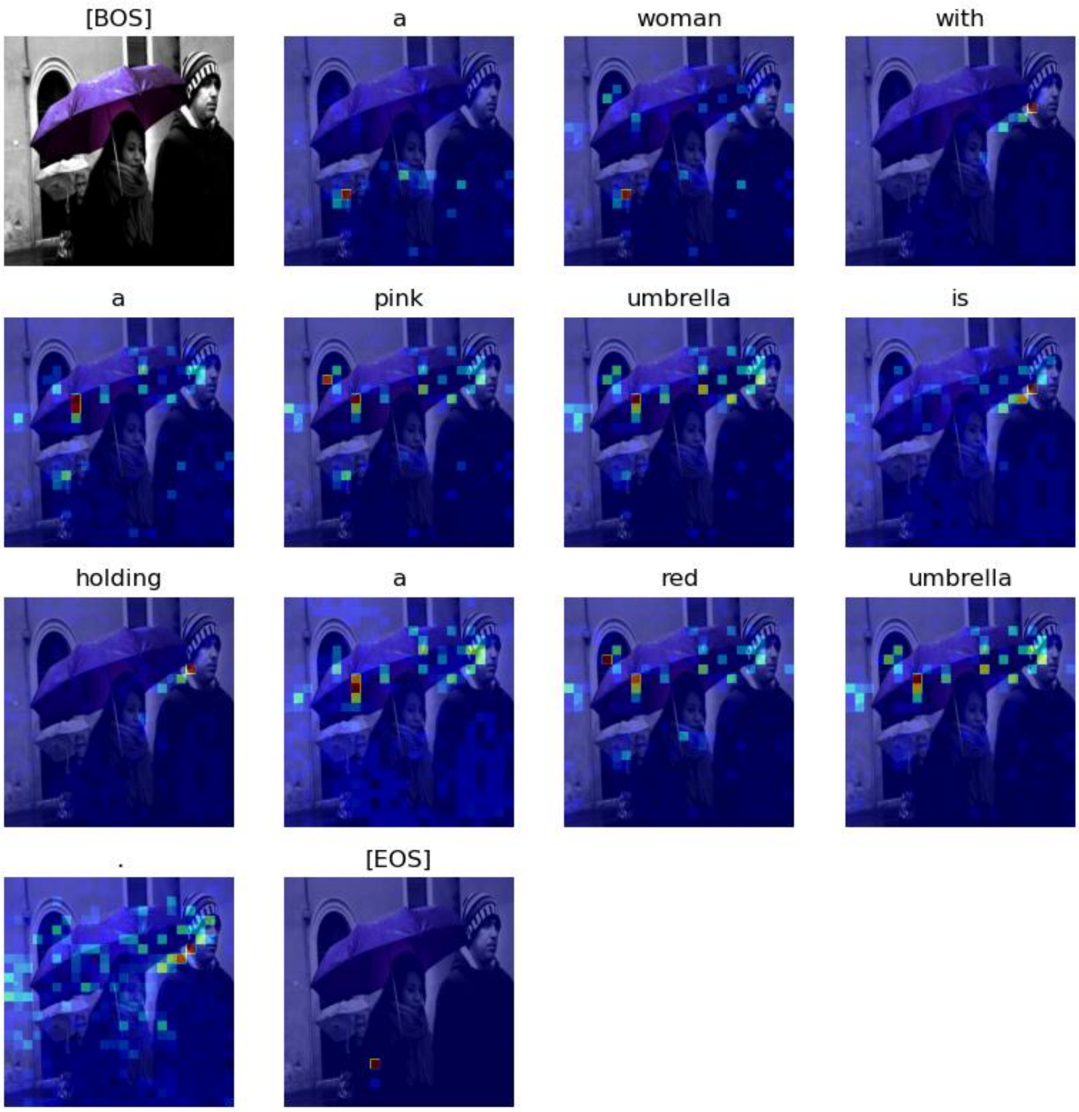
	CIDEr	CLIPScore
w/o freezing encoder	0.064	0.430
w/o label smoothing	0.789	0.686
增加 decoder 參數 (layer -> 6, feedforward-> 2048)	0.883	0.717

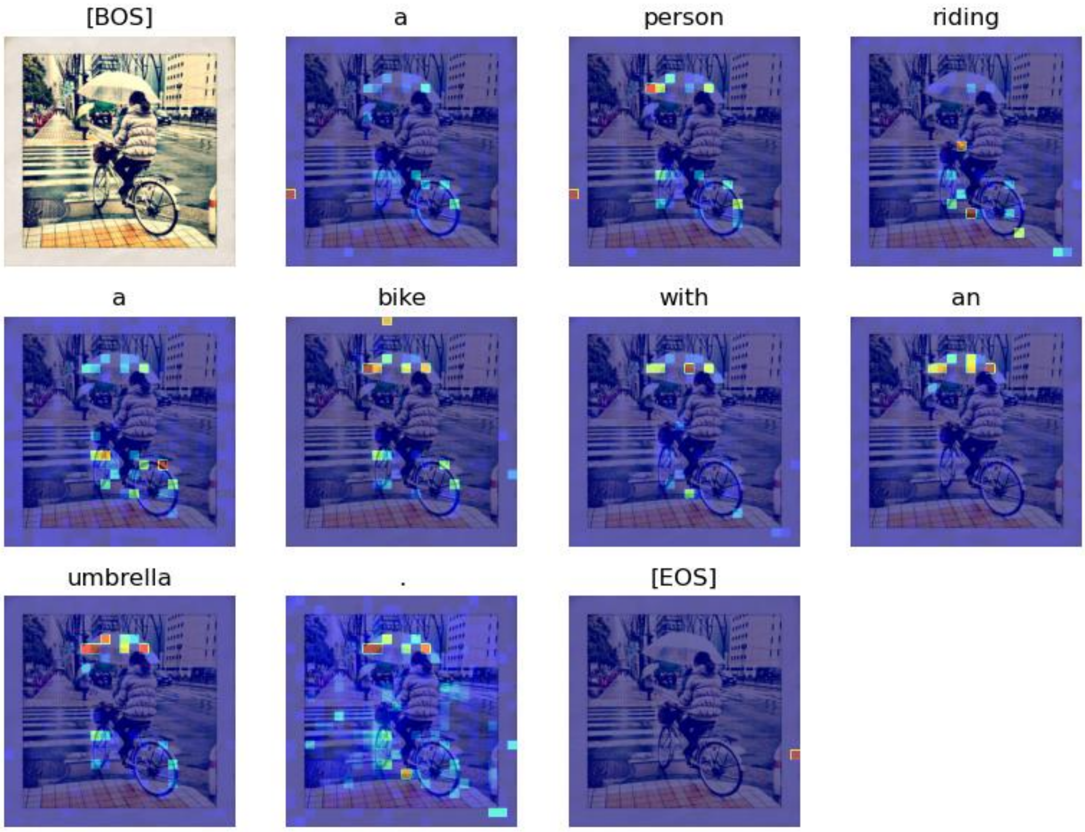
Problem 3.

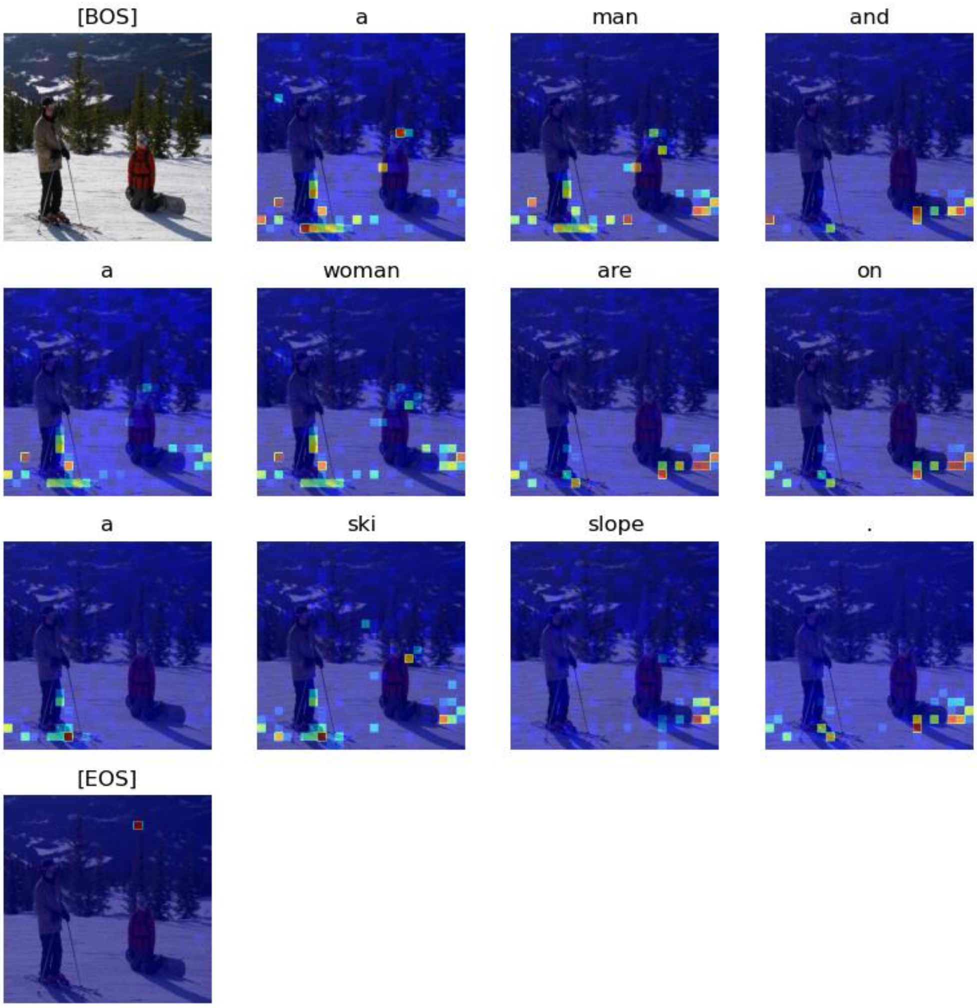
1.


[BOS]
a
small
sheep
is
standing
in
a
field
of
grass
.
[EOS]

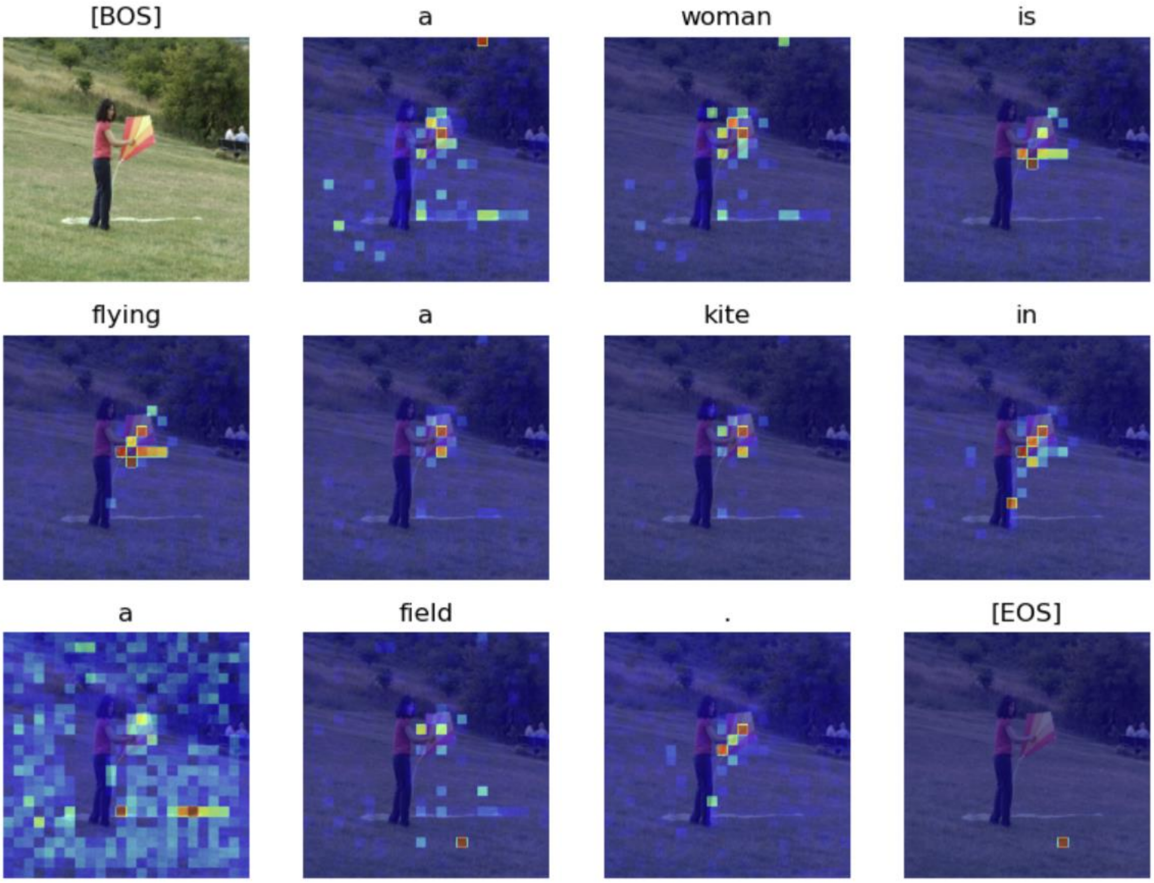

[BOS]
a
young
girl
holding
a
slice
of
pizza
.
[EOS]


[BOS]
a
woman
with
a
pink
umbrella
is
holding
a
red
umbrella
.
[EOS]

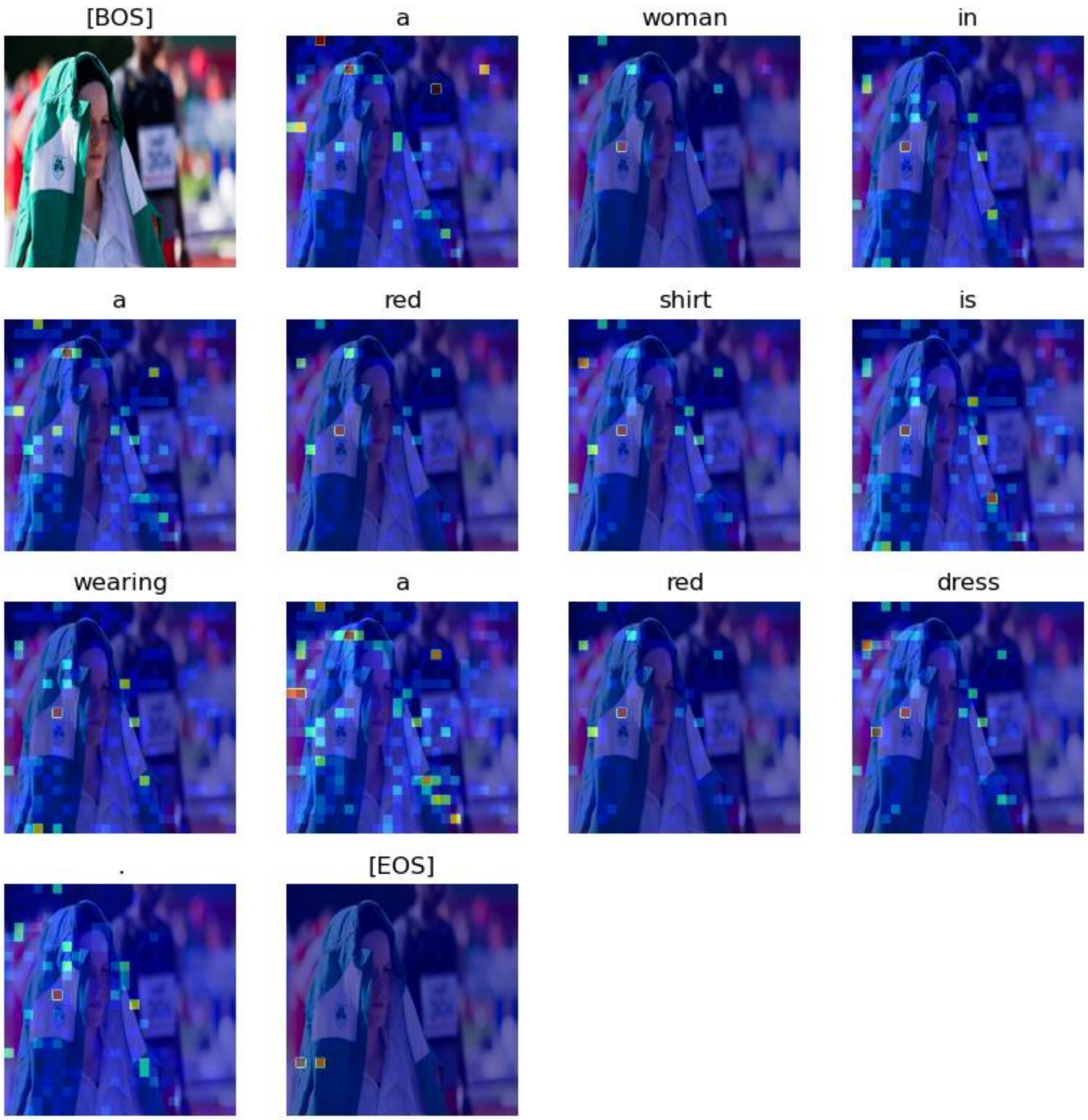

[BOS]
a
person
riding
a
bike
with
an
umbrella
.
[EOS]



2. 最高為 000000179758.jpg，其 CLIPScore = 0.997



最低為 4927180699.jpg，其 CLIPScore = 0.369



3. 有些圖片的 Attention map 在其關鍵詞(sheep, pizza…)上有反映出較大的 attention 值，也能夠有合理解釋，但也有許多的詞未必都能從 Attention map 上看出端倪(a, with,句點)，而且相鄰詞的 attention map 會較接近，且容易都注意在同一地方，這可能說明 model 還有許多進步空間