Problem1

B09901104 翁瑋杉

1. Methods analysis

- Clip is trained on a wide variety of images with a wide variety of natural language supervision that’s abundantly available on the internet .
- By not directly optimizing for the benchmark, Clip becomes much more representative.

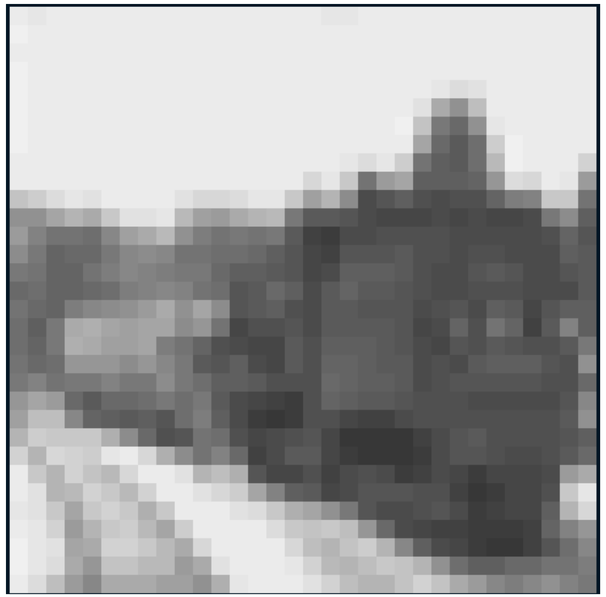
2. Prompt-text analysis

- “This is a photo of {object}”: 0.6092
- “This is a {object} image.” : 0.6788
- “No {object}, no score.”: 0.5504

3.Quantitative analysis

```
correct label: bus
0.664 bus
0.3137 streetcar
0.003431 bicycle
0.001955 pine_tree
0.001621 clock
```

```
correct label: kangaroo
0.2644 kangaroo
0.1964 camel
0.1085 elephant
0.05542 otter
0.04453 shrew
```



```
correct label: streetcar
0.8833 streetcar
0.04004 bus
0.008934 pine_tree
0.00614 dinosaur
0.005337 boy
```

Problem2

B09901104 翁瑋杉

1. Report your best setting and its corresponding CIDEr & CLIPScore on the validation data. (TA will reproduce this result) (2.5%)

```
config = {
    "seed": 1314520,
    "max_len": 55,
    "batch_size": 32,
    "num_epochs": 100,
    "val_interval": 3,
    "pad_id": 0,
    "vocab_size": 18202,
    "d_model": 768,  # 768
    "dec_ff_dim": 2048,
    "dec_n_layers": 6,
    "dec_n_heads": 12,  # 1
    "dropout": 0.1,
    "max_norm": 0.1,
}
```

```
class Encoder(nn.Module):
    def __init__(self, modelname="vit_base_patch32_224"):
        super(Encoder, self).__init__()
        self.vit = timm.create_model(modelname, pretrained=True)

    def forward(self, x):
        x = self.vit.forward_features(x)

        return x
```

Score: CLIPScore: 0.744 | CIDEr: 0.8844

Report other 3 different attempts (e.g. pretrain or not, model architecture, freezing layers, decoding strategy, etc.) and their corresponding CIDEr & CLIPScore. (7.5%, each setting for 2.5%)

- Decoding strategy:
  - Regular method: CLIP score: 0.7617 | CIDEr score 0.8556
  - Beam search: CLIPScore: 0.744 | CIDEr: 0.8844
  - By using beam search, CLIPScore has decline but CIDEr has increased.

- Freezing encoder or not:
  - Freezing encoder: CLIP score: 0.7617 | CIDEr score 0.8556
  - Not freezing encoder(train both decoder and pretrained encoder): CLIP score: 0.4937 | CIDEr score 0.1426
  - It implies that it is not feasible to train encoder and decoder simultaneously.

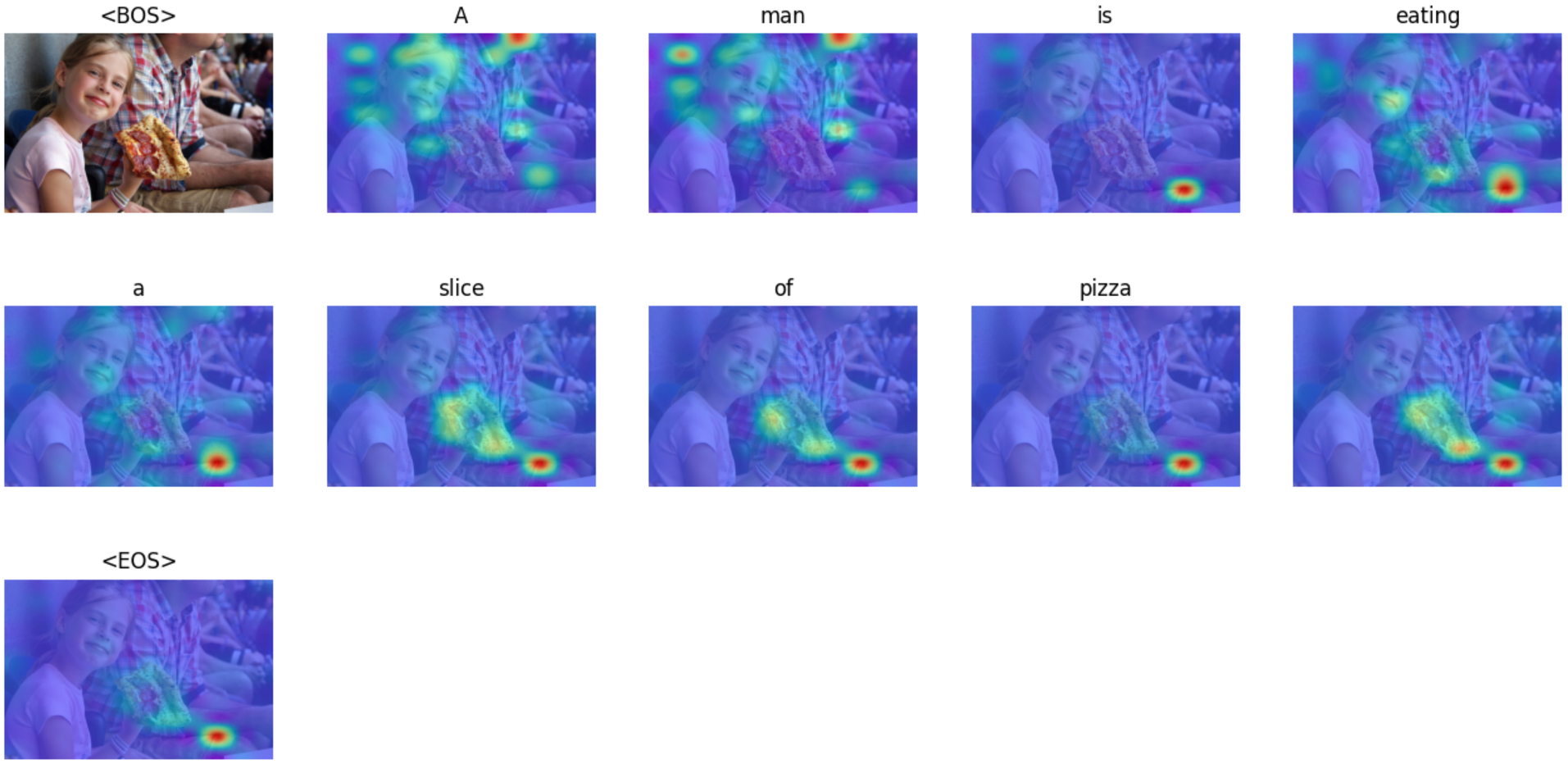
- Different ViT (both decoding by beam search)
  - vit_large_r50_s32_384: CLIP score: 0.7251 | CIDEr score 0.8852
  - vit_base_patch32_224: CLIPScore: 0.744 | CIDEr: 0.8844

Problem3

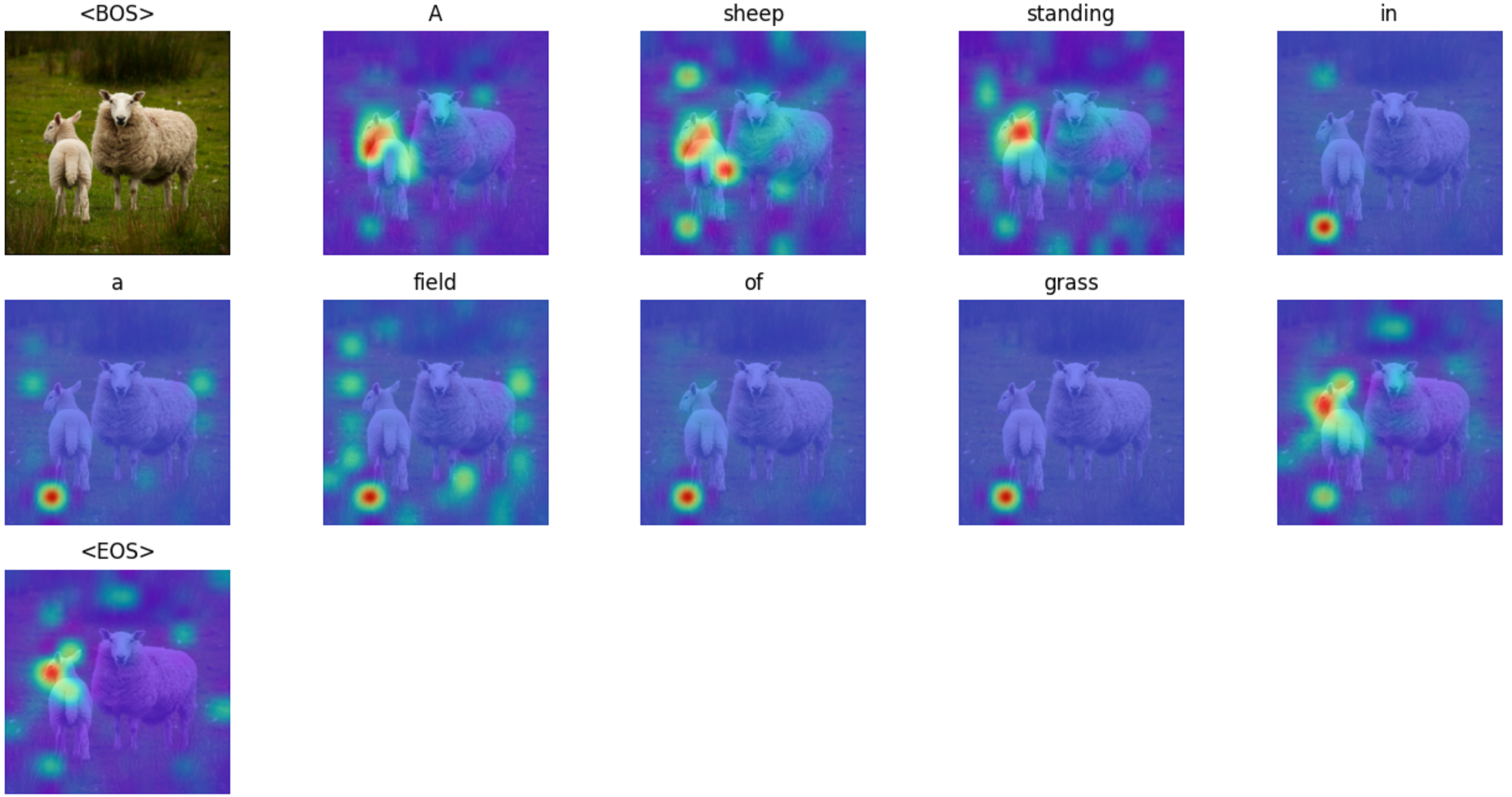
B09901104 翁瑋杉

1.Attention map visualization

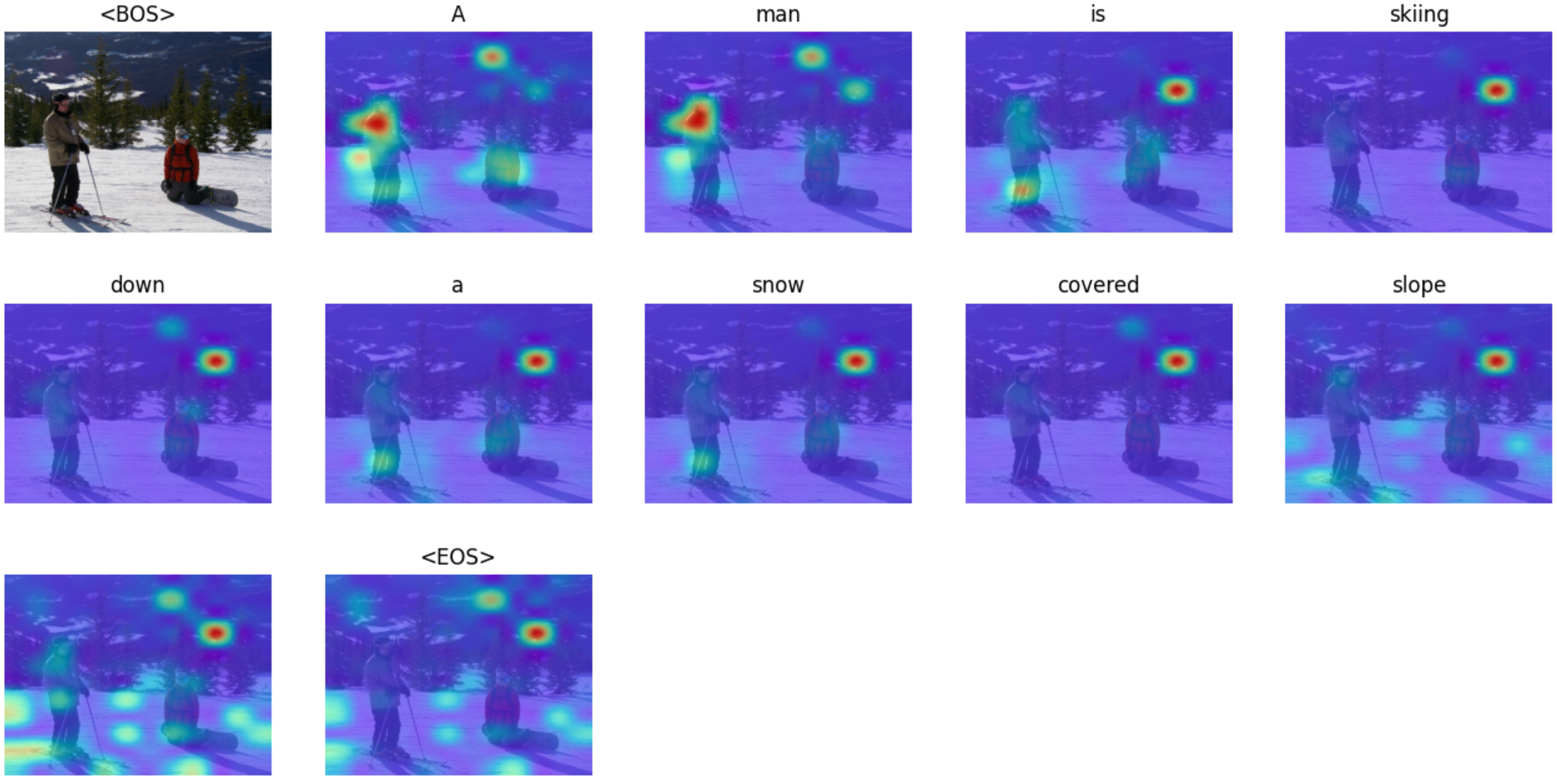
A man is eating a slice of pizza .



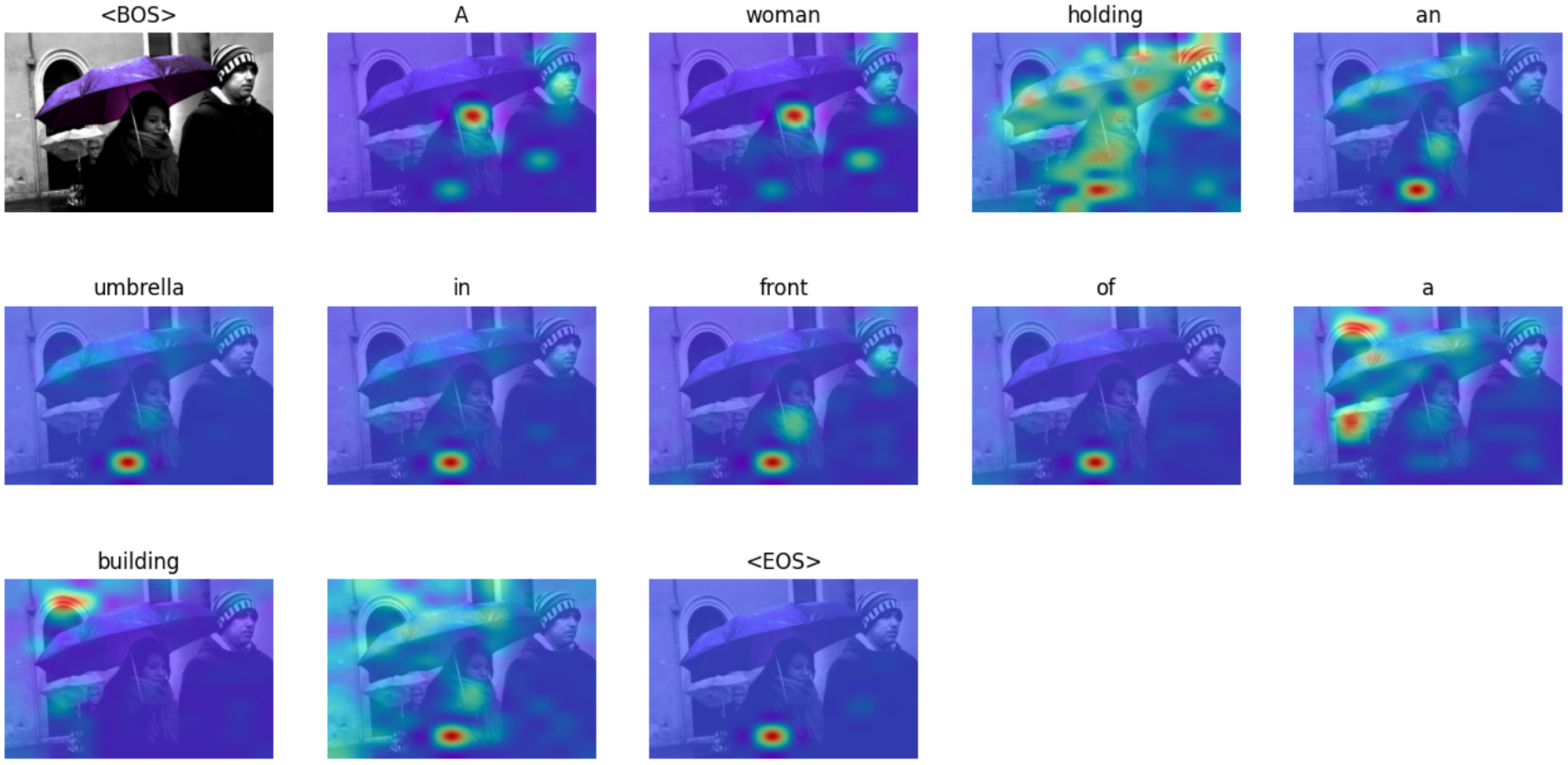
A sheep standing in a field of grass .



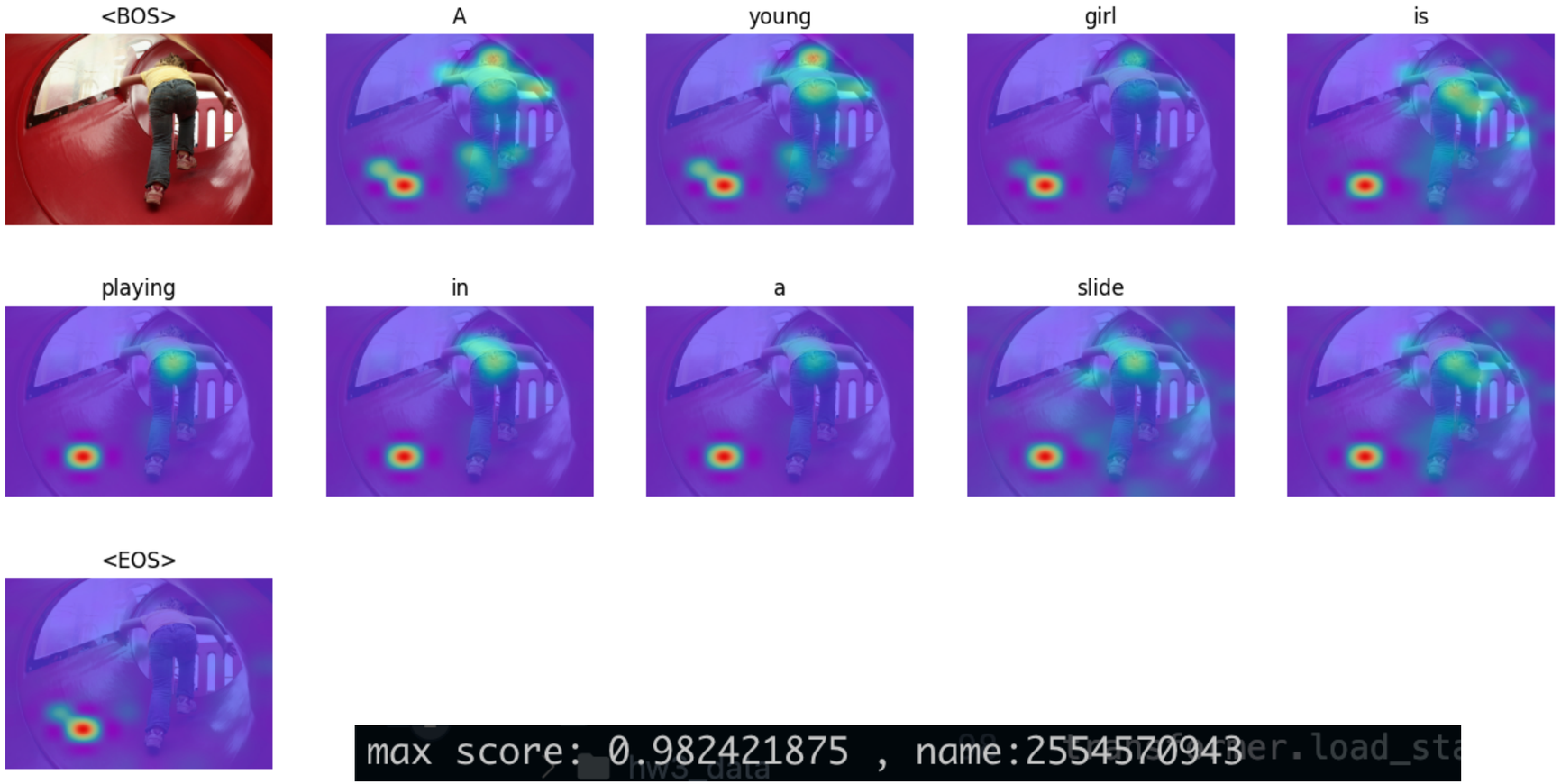
A man is skiing down a snow covered slope .



A woman holding an umbrella in front of a building .

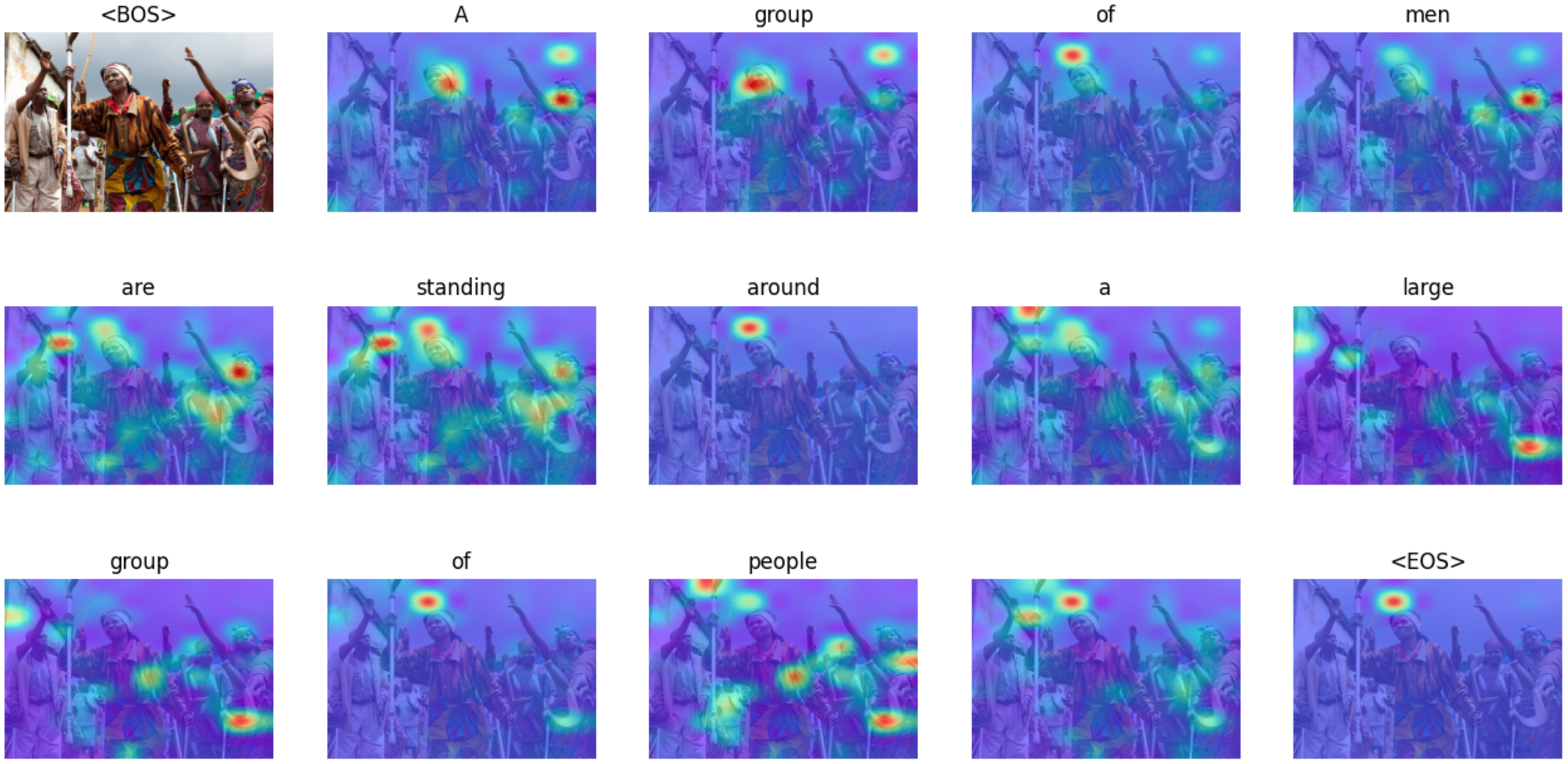


According to **CLIPScore**, you need to visualize:
1.top-1 and last-1 image-caption pairs
2.its corresponding CLIPScore in the validation dataset of problem 2. (5%)
Analyze the predicted captions and the attention maps for each word according to the previous question. Is the caption reasonable? Does the attended region reflect the corresponding word in the caption? (5%)




```
max score: 0.982421875 , name:2554570943
```

- I think the caption is pretty reasonable and highly corresponds the image. Beside the dot at bottom left corner, the heat map basically does reflect the corresponding word in the caption (for example, “young”, “girl” and “slide”).




```
min score: 0.41748046875 , name:5109882423[]
```

- I think the caption is pretty lame. However, if we look at the “men” , “standing” and “people” heat map, they do correspond to the predicted words.
- Overall, I think transformer is explainable to a certain degree, but there’re still something in the black box that we can hardly understand or interpret.