

# DLCV HW3 Report

R11921008 羅恩至

## Problem1

**(3%) Methods analysis. Previous methods (e.g. VGG and ResNet) are good at one task and one task only, and requires significant efforts to adapt to a new task. Please explain why CLIP could achieve competitive zero-shot performance on a great variety of image classification datasets.**

Traditional CNN methods need training label for each image, sometimes if the images quality is not good enough or the data is mislabeled, or the domain between training set and testing set is large, this will weaken the performance of the model because the performance of CNN models is highly influenced by the quality and quantity of the training data.

Different with traditional CNN, CLIP model learns the similarity between the texts and the images, and compares the similarity between the two. Images and text are encoded with image encoder and text encoder to get the image and text features, and by comparing the cosine similarity and output the top few labels with highest probability, CLIP can get the output label of the input image. So CLIP can achieve good performance without labeled input data.

**(6%) Prompt-text analysis. Please compare and discuss the performances of your model with the following three prompt templates: (1) “This is a photo of {object}”, (2) “This is a {object} image.” (3) “No {object}, no score.”**

Prompts	Validation Accuracy
“This is a photo of {object}”	0.6076
“This is a {object} image.”	0.6820
“No {object}, no score.”	0.5628

The prompt text analysis is tested on model Vit-B/32. From the performance result of three different prompts, it shows that prompt “This is a {object} image.” has the best performance with accuracy 0.6820, and prompt “No {object}, no score.” has the lowest performance with accuracy only 0.5628. From the experiment result, the performance of model is surely related to the prompts. And to my perspective, the prompts with words that can form a complete normal sentence or contain words that


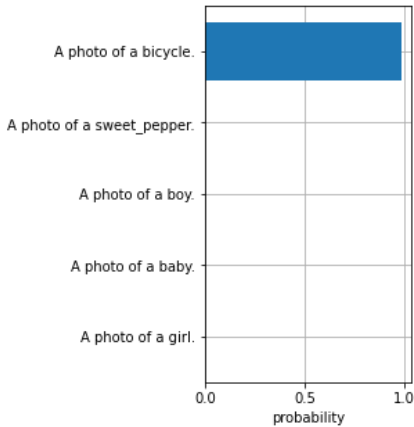
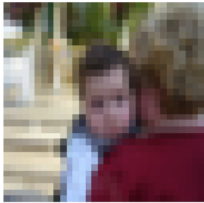
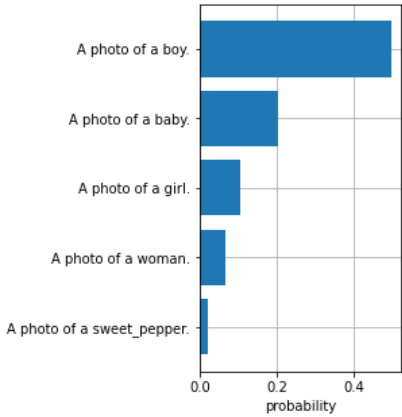
are able to describe the picture, like prompt “This is a photo of {object}”, prompt “This is a photo of {object}” and inference prompt “A photo of a {object}.” all have a good performance, and the prompts that form incomplete sentences, abnormal grammar or contain not such related words to well describing the picture would get lower performances, like prompt “No {object}, no score.” with a performance only 0.5628.

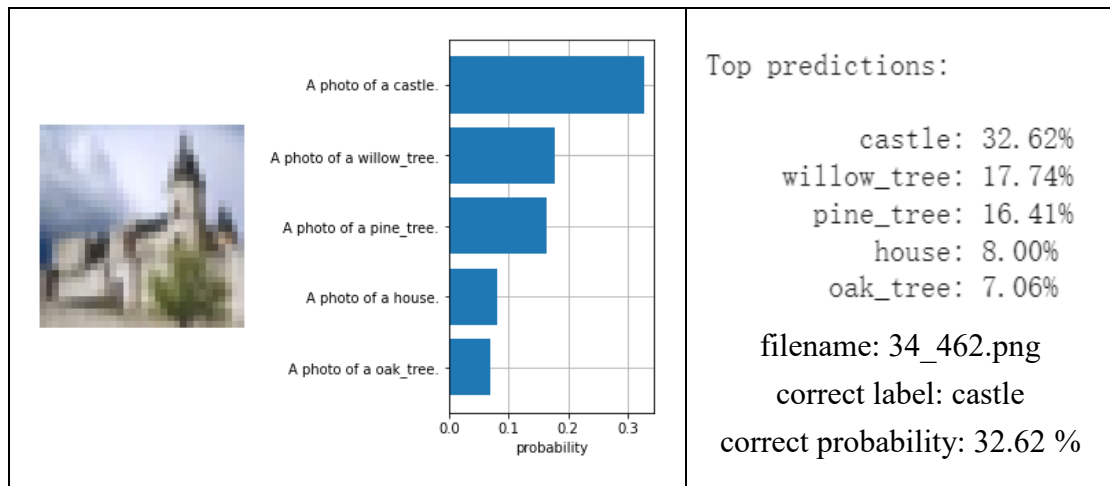
**(6%) Quantitative analysis. Please sample three images from the validation dataset and then visualize the probability of the top-5 similarity scores.**

Model: Vit-B/32

Prompt: “A photo of a {object}.”

Results:

 	<p>Top predictions:</p> <p>bicycle: 98.34% sweet_pepper: 0.46% boy: 0.11% baby: 0.10% girl: 0.09%</p> <p>filename: 0_477.png correct label: bicycle correct probability: 98.34 %</p>
 	<p>Top predictions:</p> <p>boy: 49.71% baby: 20.40% girl: 10.75% woman: 6.84% sweet_pepper: 2.29%</p> <p>filename: 10_463.png correct label: baby correct probability: 20.40 %</p>



## **Problem 2**

Reference: CATR: Image Captioning with Transformers

<https://github.com/saahiluppal/catr>

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

Strategy	Settings
Image size	384*384
Position embedding	sine
Number of layers of transformer encoder and decoder	6
Hidden dimension	320
Number of heads	10
Max word embedding length	160
Learning rate	1e-05
Decoding strategy	Greedy search

Score:

CIDEr	CLIPScore
0.5943	0.6501

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

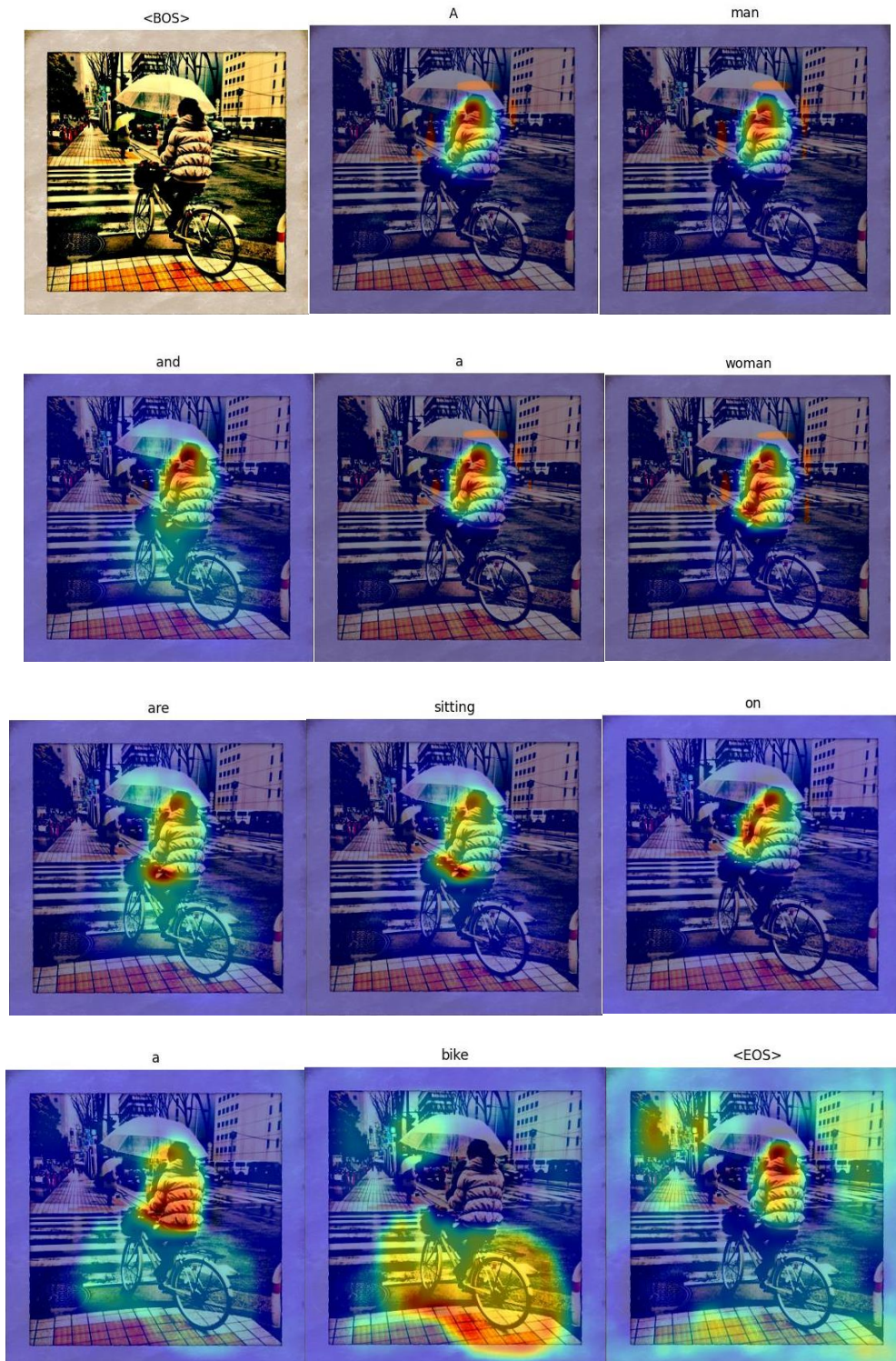
Method	CIDEr	CLIPScore
All train from scratch	0.5816	0.6433
Train with proposed pretrained encoder and random initialize decoder weights	0.191	0.544
Train with timm vit_large_patch16_224 pretrained model as encoder	0.191	0.447

### Problem 3

(10%, each image for 2%) TA will give you five test images ([p3\_data/images/]), and please visualize the predicted caption and the corresponding series of attention maps in your report with the following template.

1. bike.jpg

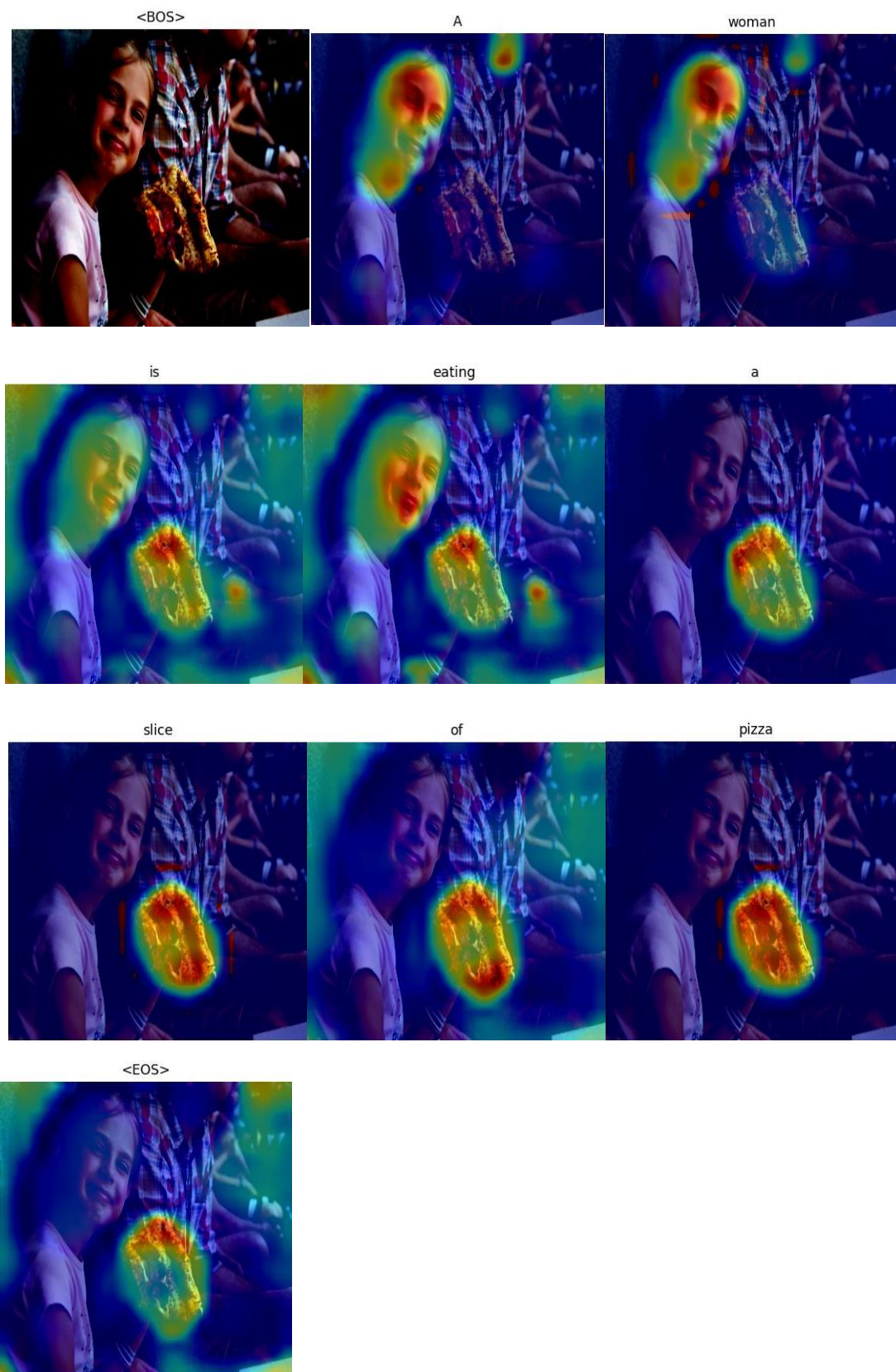
Caption: A man and a woman are sitting on a bike





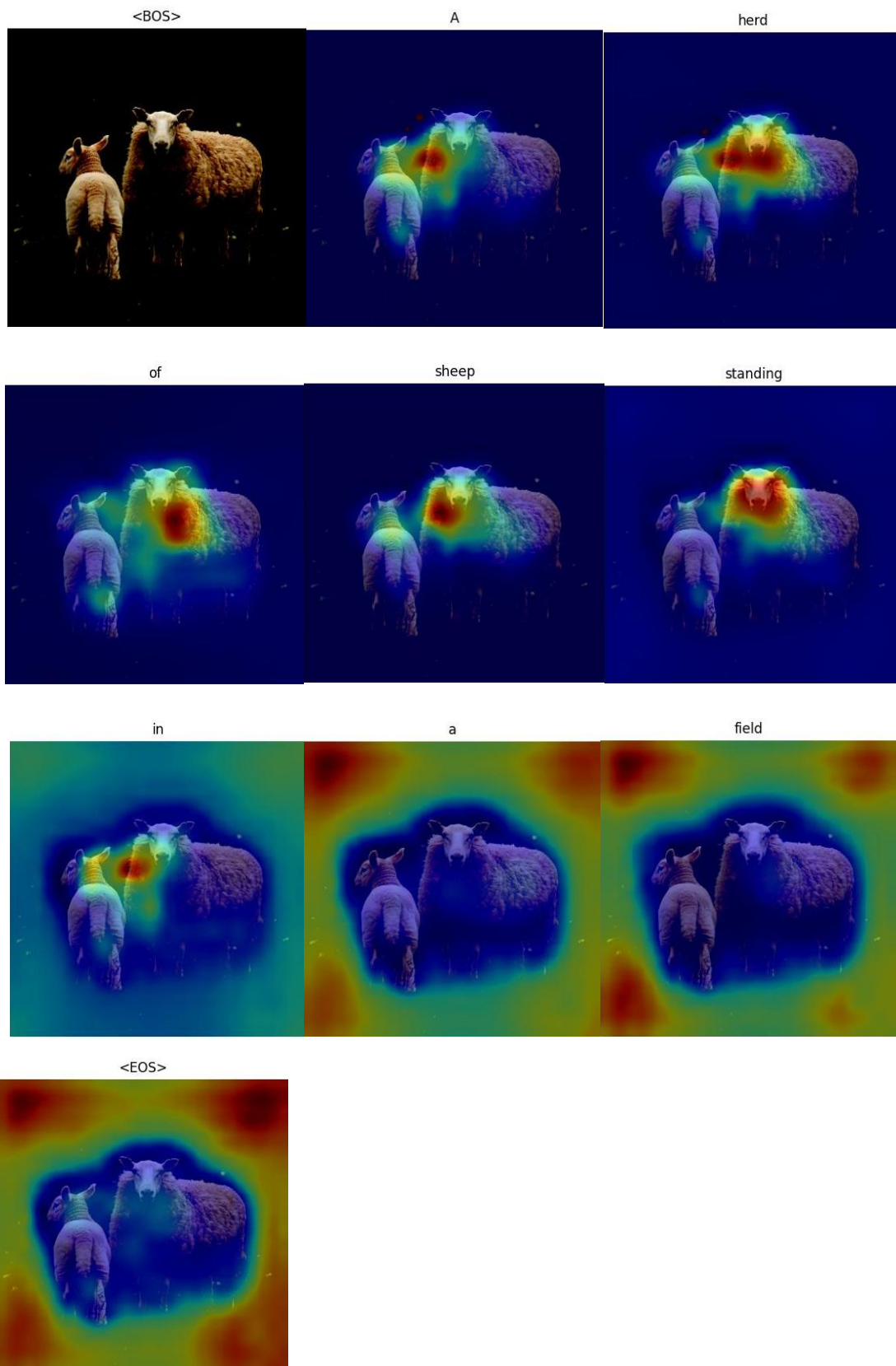
2. girl.jpg

Caption: A woman is eating a slice of pizza



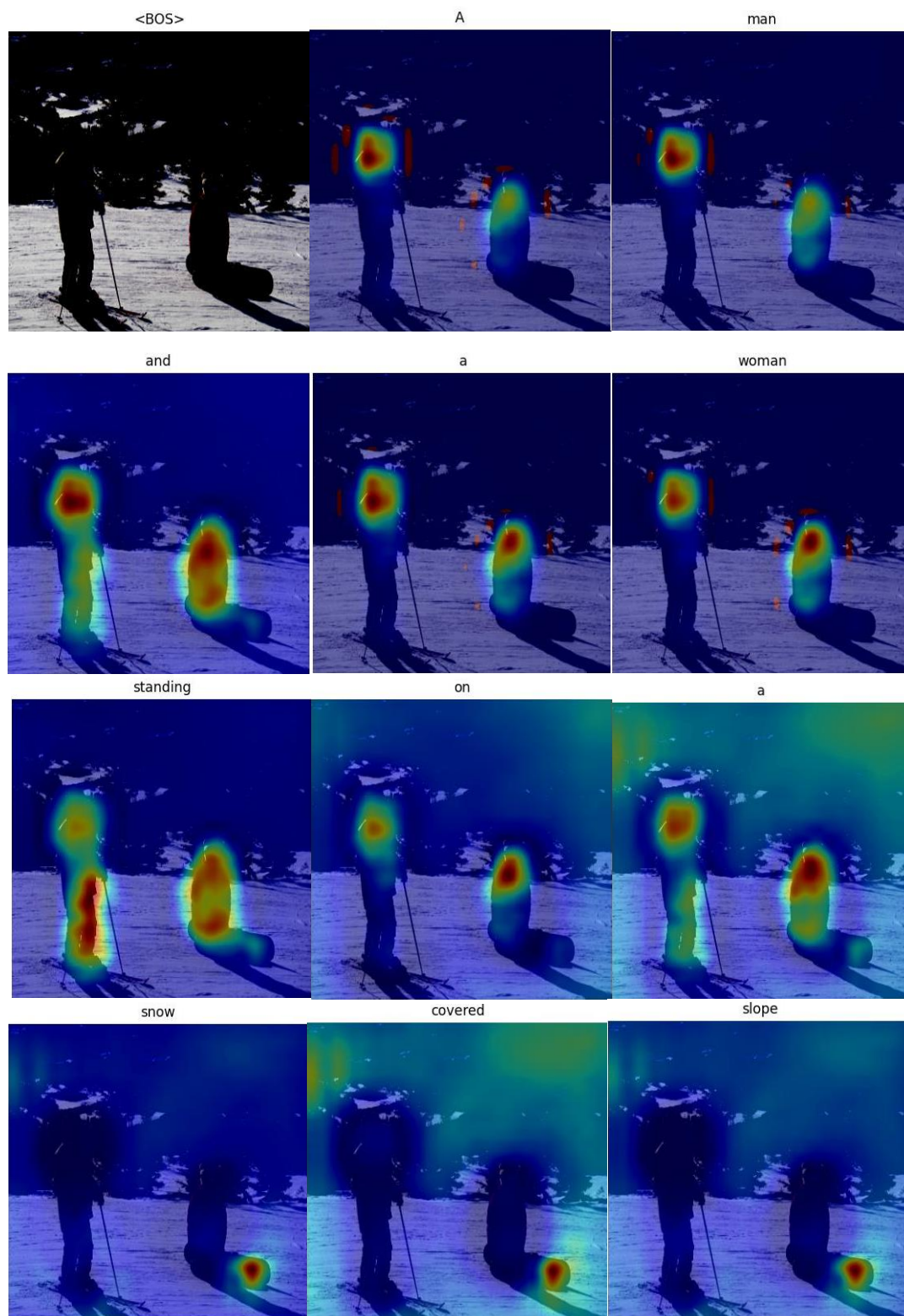
### 3. sheep.jpg

Caption: A herd of sheep standing in a field



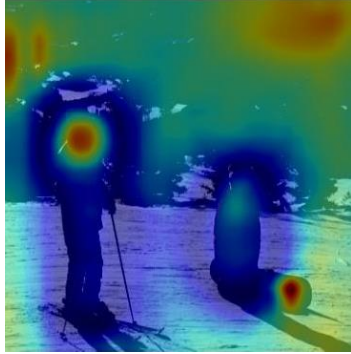
4. ski.jpg

Caption: A man and a woman standing on a snow covered slope





<EOS>



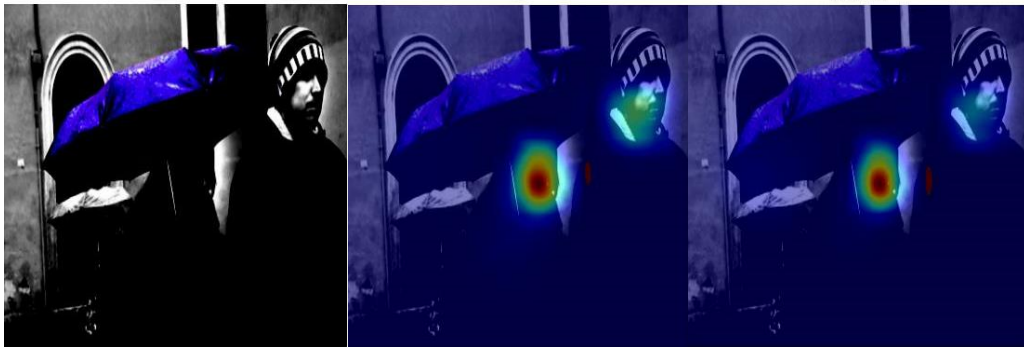
5. umbrella.jpg

Caption: A woman with a blue shirt and a blue shirt

<BOS>

A

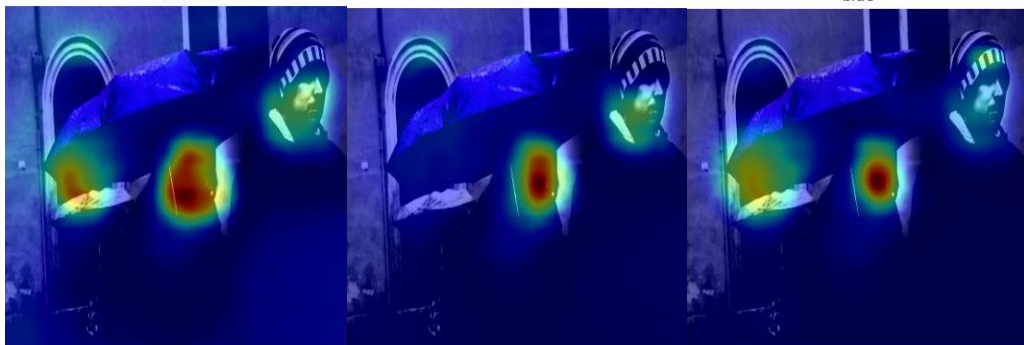
woman



with

a

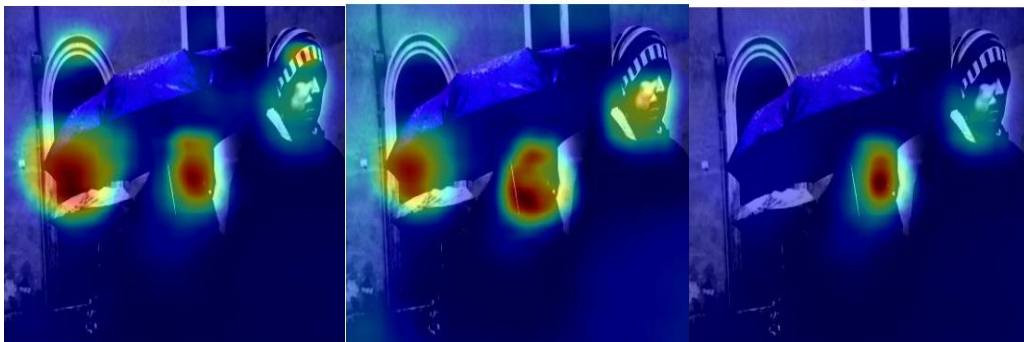
blue

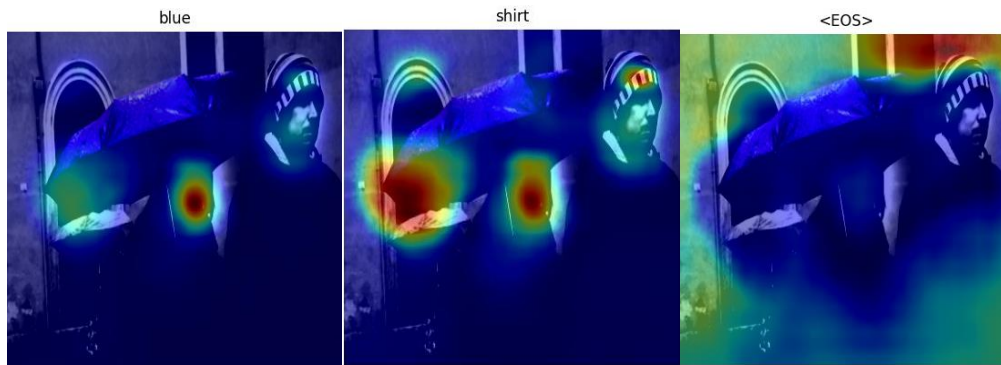


shirt

and

a





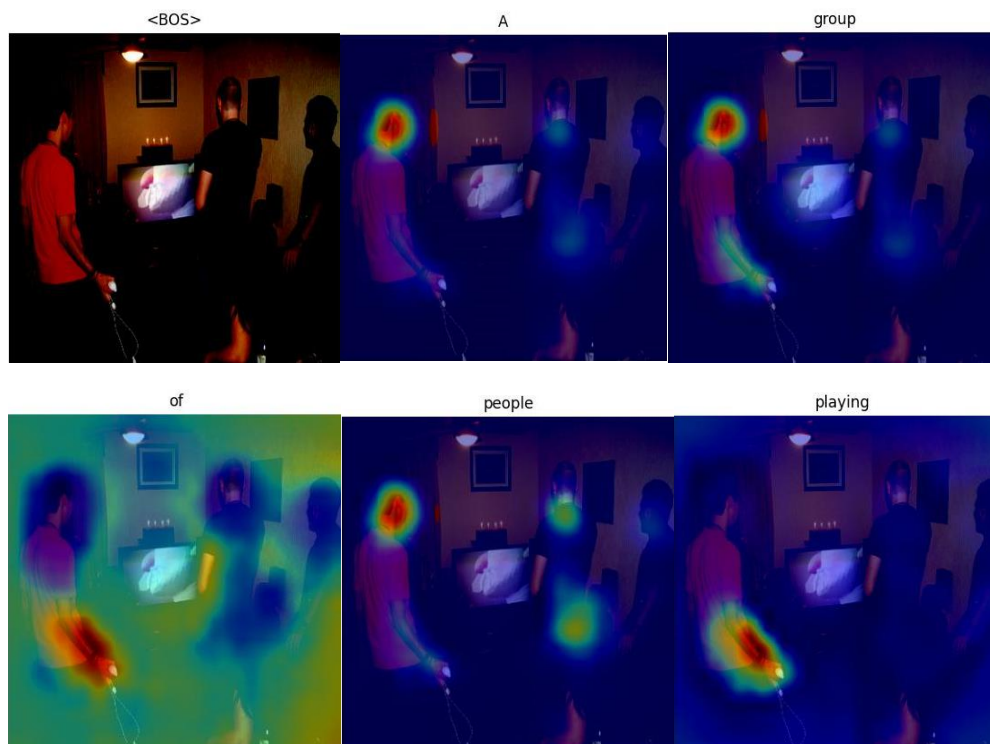
**(5%) According to CLIPScore, you need to visualize: top-1 and last-1 image-caption pairs its corresponding CLIPScore in the validation dataset of problem 2.**

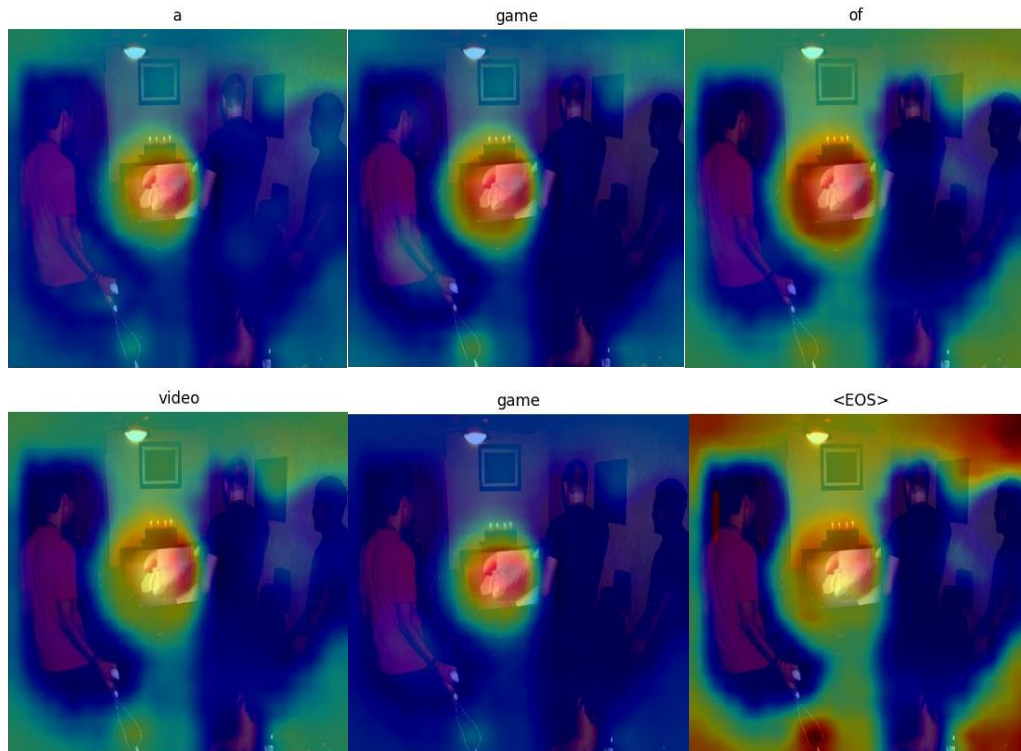
(1) Top-1 image-pairs

Image file: 000000406755.jpg

CLIPScore: 0.9369

Caption: A group of people playing a game of video game



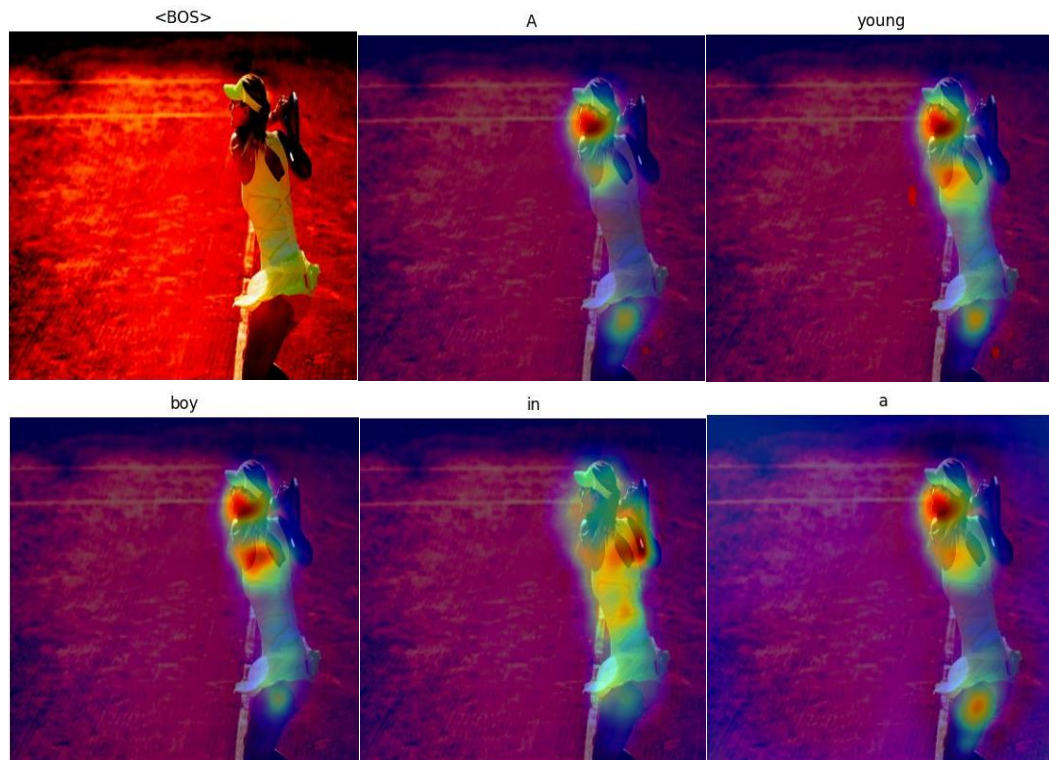


(2) Last-1 image-pairs

Image file: 000000084157.jpg

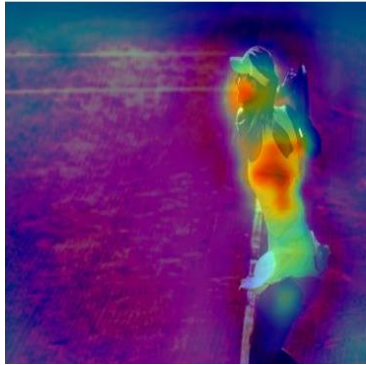
CLIPScore: 0.2484

Caption: A young boy in a blue and blue shirt is jumping into a pool

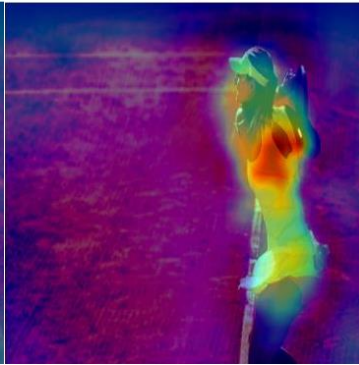




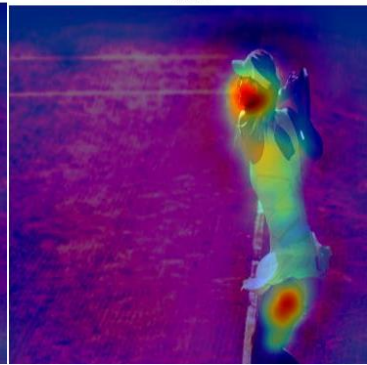
blue



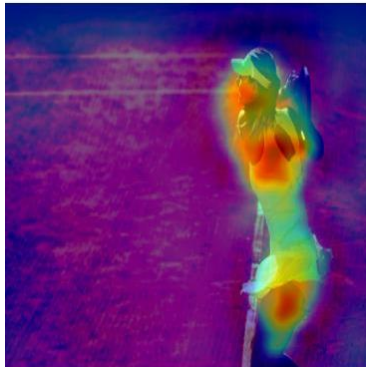
and



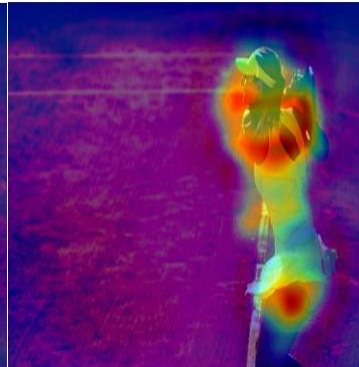
blue



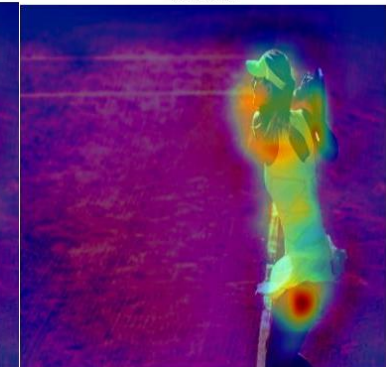
shirt



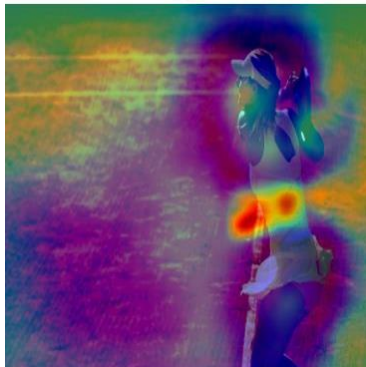
is



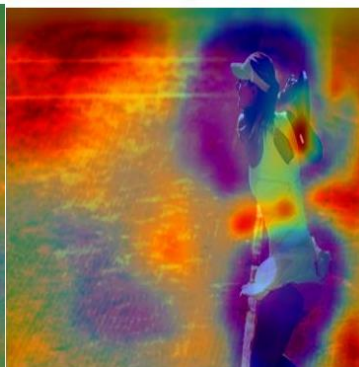
jumping



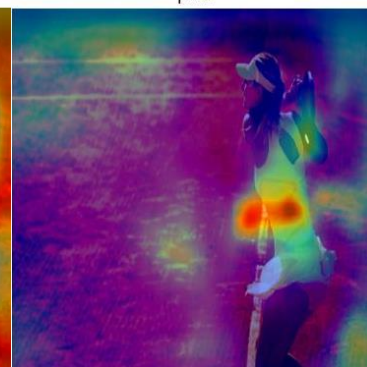
into



a



pool



<EOS>





**(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%)**

For the caption and visualization attention map results of these 7 images from previous questions, I think the caption result is sometimes not quite reasonable. Sometimes the model may not correctly distinguish the human, objects, or actions in the picture. For example, in image bike.jpg, the caption outputs “A man and a woman” while there is only a woman in the image, and in 000000084157.jpg, the model mistakes “the woman” to a “young boy”, and “playing tennis” to “jumping into a pool”. However in sheep.jpg, 000000406755.jpg, the caption result is reasonable, and other image captions are also reasonable with small mistake or neglect some information. Like ski.jpg the caption doesn’t contain “ski” inside but only “snow” in it. In overall, the captions inference from my model may sometimes mistake or neglect some information.

As of the attention map result, I think the mapping result is approximately reasonable. In overall, the nouns will mainly focus on the representing object, like in girl.jpg, the word “pizza” highly focus on the pizza in the image, and in bike.jpg the word “bike” focus on the bike. The attention maps of verbs are referring to the part that taking the action, like in 000000084157.jpg the word “playing” highly focus on the hand with a joystick of the man, and in girl.jpg the word “eating” highly focus on the girl’s mouth. As of prepositions, like “in”, “on” or “of” or the end token “<EOS>”, the model seems to have difficulty to find which part to focus on, so the attention map of these words is sparse.