DLCV HW3 REPORT

r12942046 薛仁豪

Problem 1

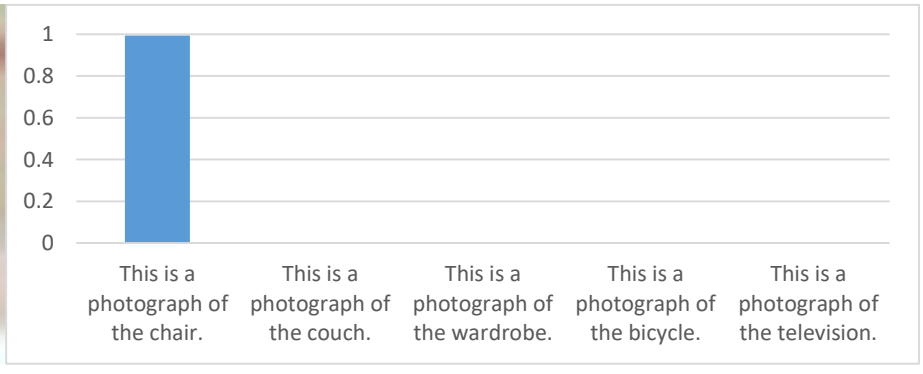
1. Methods Analysis

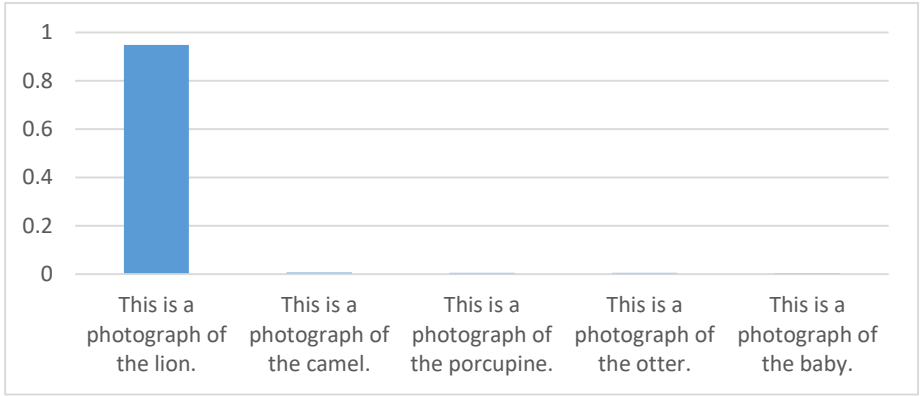
CLIP is trained on data accompanied by textual descriptions, employing contrastive learning. As a result, it possesses the capability to effectively process a greater amount of information.

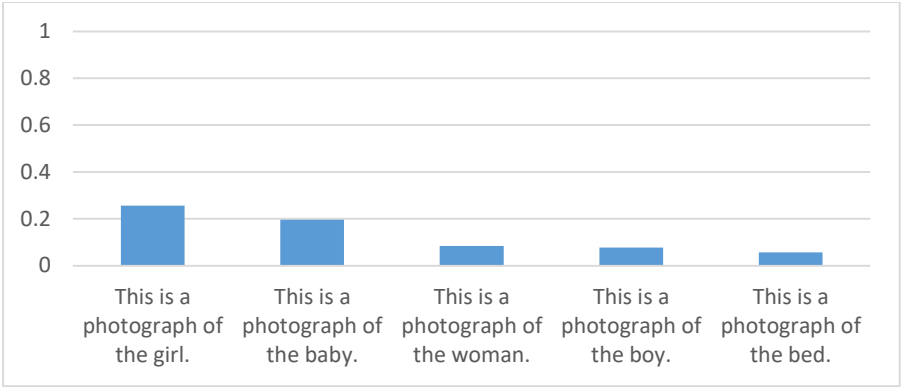
2. Prompt-text Analysis

	Accuracy
“This is a photo of {object}”	**0.6080**
“This is not a photo of {object}”	**0.6532**
“No {object}, no score.”	**0.5636**

1. Quantitative Analysis







Problem 2-1

1. Best Settings

The encoder is vit_huge_patch14_clip_336, with adapter added in every blocks of the decoder

	CIDEr	CLIPScore
Best	**0.8322**	**0.7081**

2. Different Attempts of PEFT

	CIDEr	CLIPScore
Adapter	**0.8322**	**0.7081**
Prefix Tuning	**X**	**X**
Lora	**0.4113**	**0.5206**

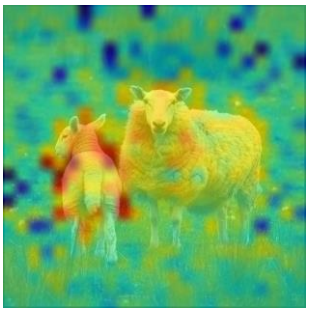
I don’t know why but my LoRa disrupts

Problem 2-2

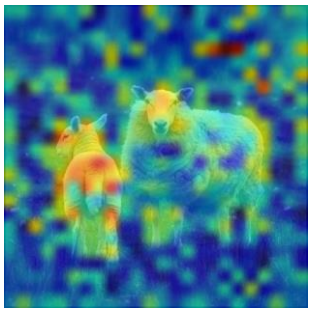
1. Attention Maps

SHEEP: a herd of sheep standing together in a grassy area.

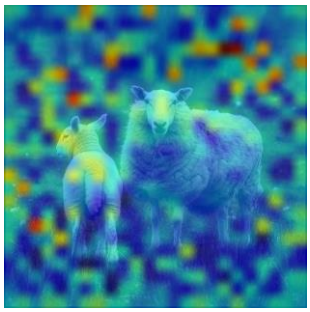
a



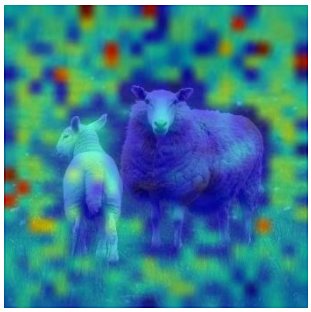
herd



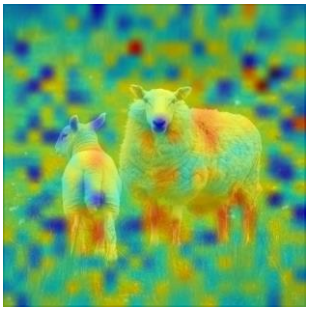
of



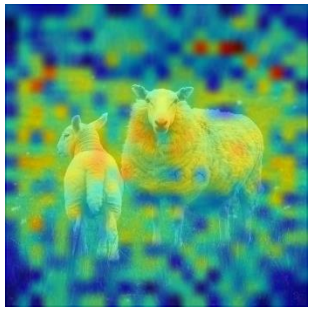
sheep



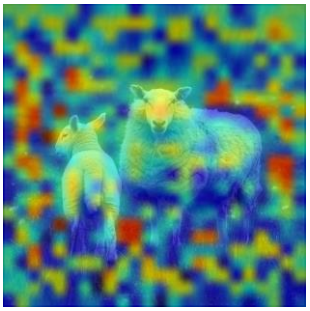
standing



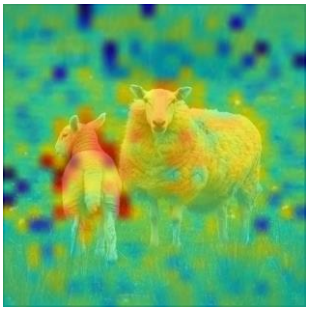
together



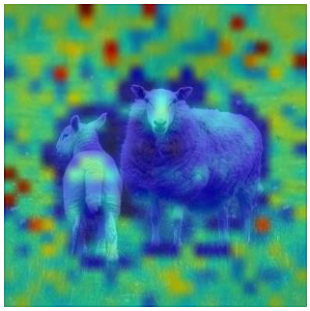
in



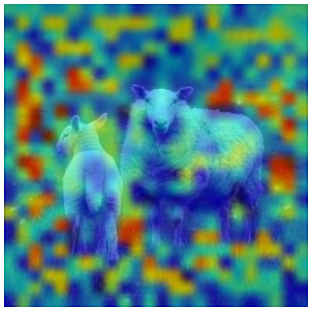
a



grassy

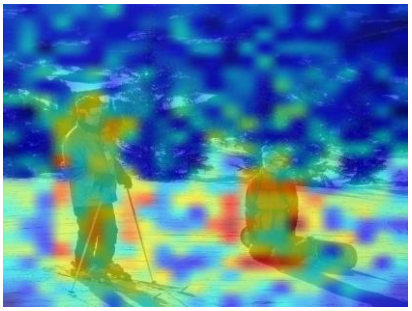


area.

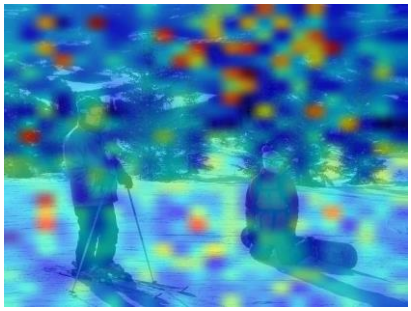


SKI: a group of people skiing down a mountain.

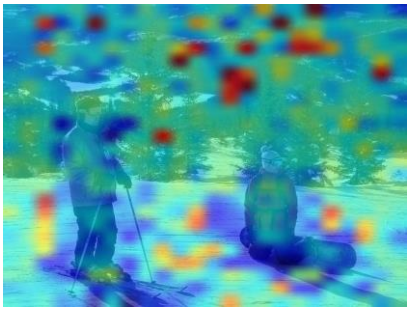
a



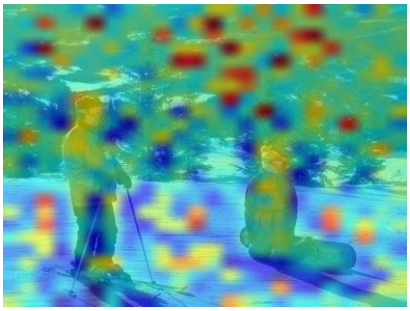
group



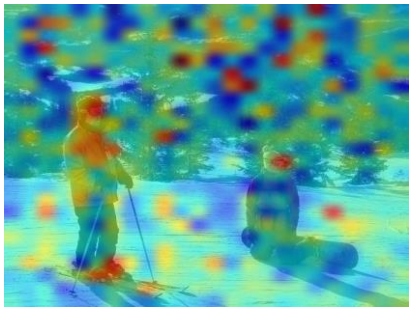
of



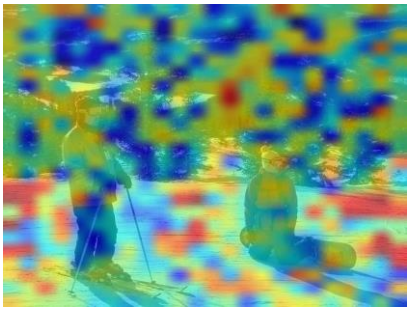
people



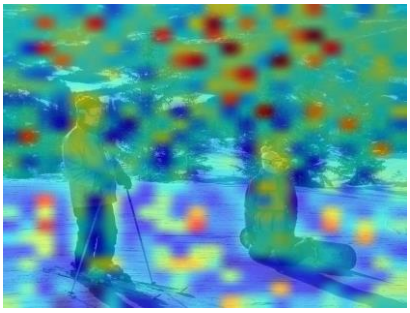
skiing



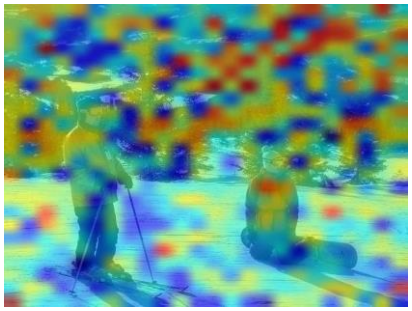
down



a



mountain.

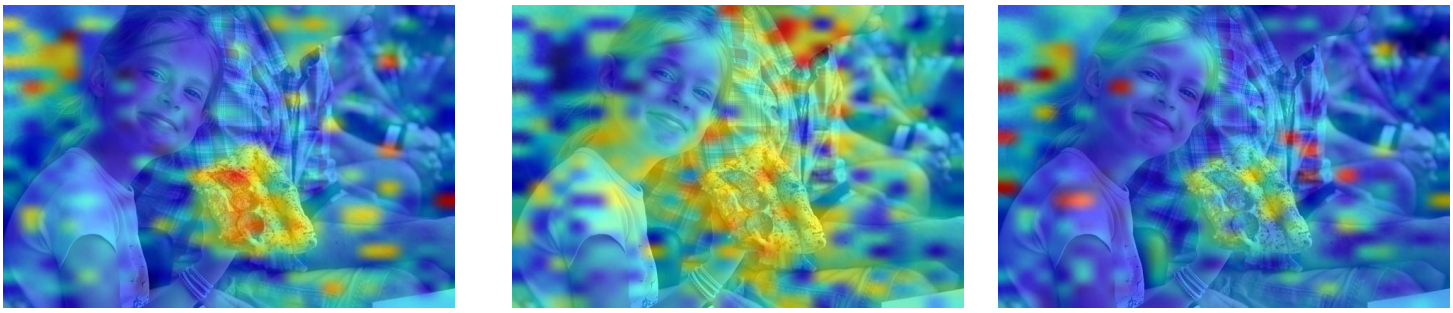


GIRL: a young girl sitting down with her father while holding a pizza.

a **young** **girl**

sitting **down** **with**

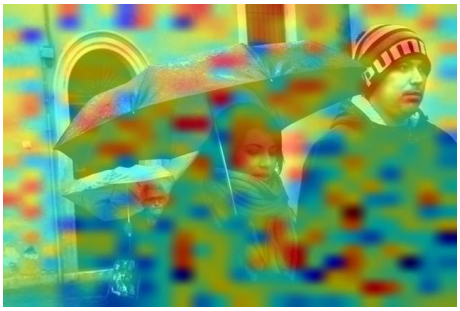
her **father** **while**



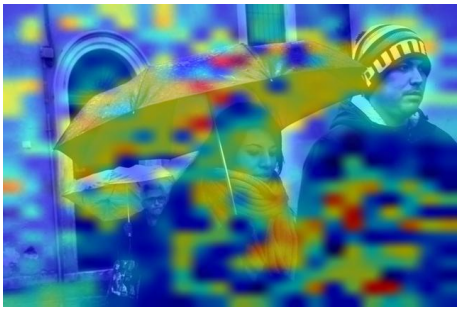
holding **a** **pizza**

Umbrella: a women with curly hair sits down with her husband.

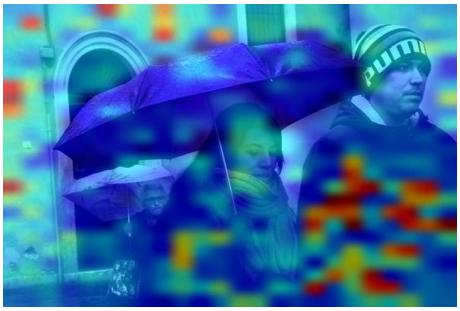
a



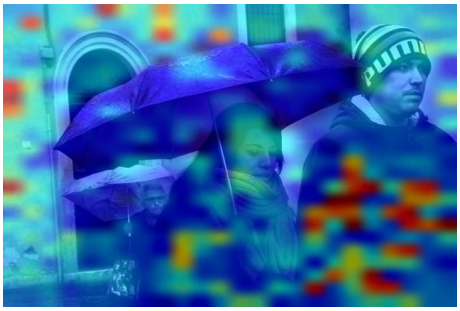
woman



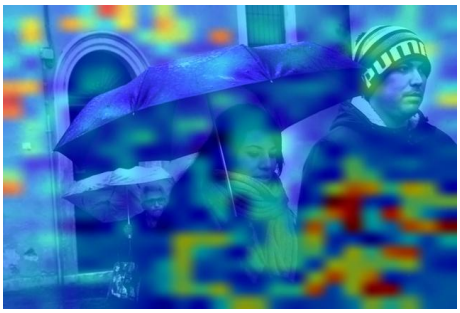
with



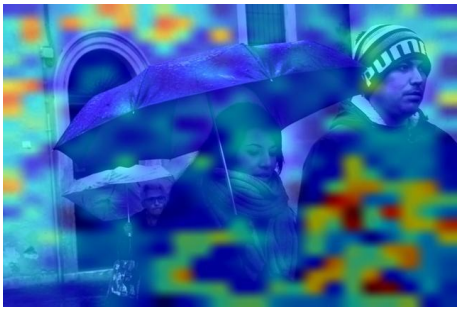
curly



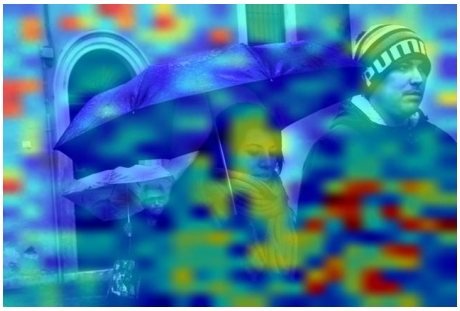
hair



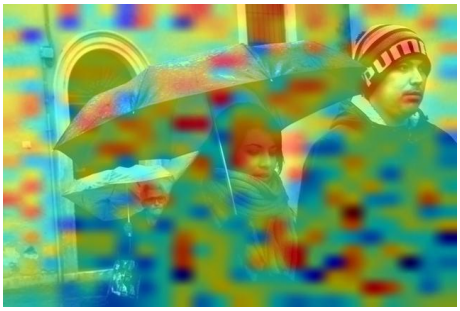
sits



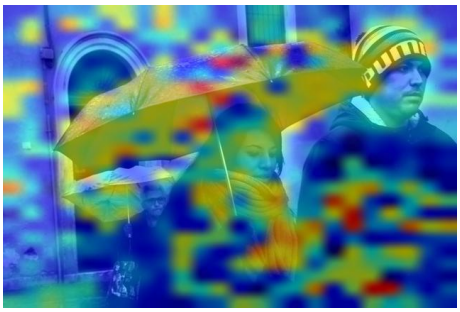
down



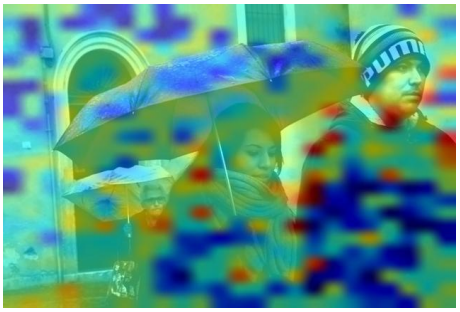
with



her



husband



Bike: a women riding a bike with her umbrella attached to her bike.

a

woman

riding

a

bike

with

her

umbrella

attached

to

her

bike.

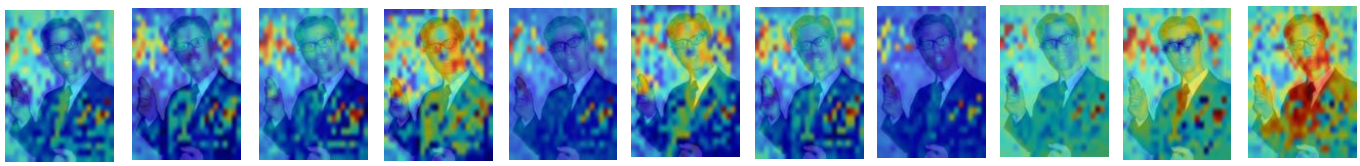
2. Visualization of Best and Worst Image

Best

a man wearing a suit and tie holding a banana. (0.98)

Worst

a woman in a red shirt holding a red hat. (0.37)



3. Analyzation

It can be observed that in the worst-case scenario, the attention head may not capture the correct features effectively.