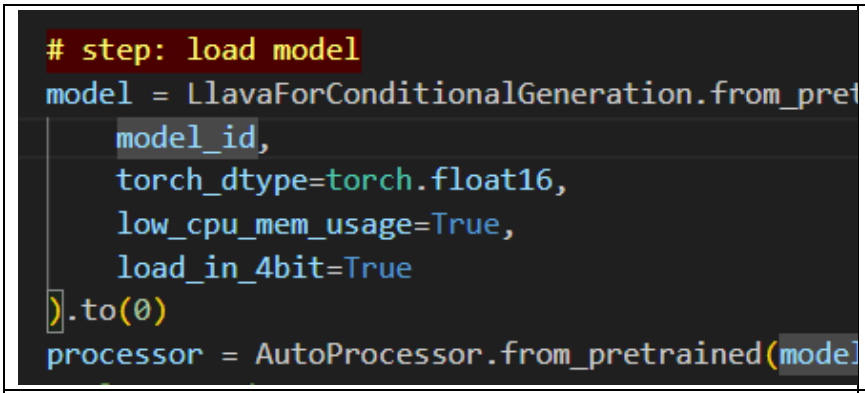
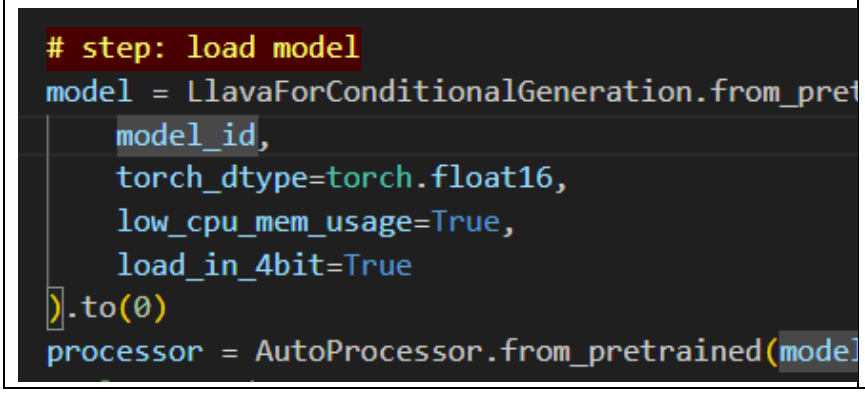
Problem 1

Paper reading(3%): Please read the paper “Visual Instruction Tuning” and briefly describe the important components (modules or techniques) of LLaVA.

- Modules:
    - **Vision Encoder**: It uses the CLIP visual encoder (ViT-L/14) to extract visual features.
    - **Projection Layer**: It maps the visual features to the language embedding space through a trainable linear projection matrix.
    - **Language Model**: It employs the Vicuna language model, which has robust instruction-following capabilities for language tasks.
- Technique
    - It utilizes language-only models like GPT-4 to generate diverse and detailed multimodal instruction-following datasets, including conversation-style questions, detailed descriptions, and complex reasoning questions, related to images.

Prompt-text analysis (6%): Please come up with two settings (different instructions or generation config). Compare and discuss their performances.

Setting	CIDEr & CLIPScore
Instruction: “Please give me a one sentence caption about the image.” Max token= 200	CIDEr: 1.1604811812238933 CLIPScore: 0.7778149414062 5

	
Instruction: “describe the image in one sentence” Max token= 200 	CIDEr: 1.0774418715436025 | CLIPScore: 0.786458740234375

Thw first one use the “caption” as the prompt and the second one use the “describe” as the prompt, the first one get better result on CIDEr, while the second one get better result on CLIPScore, which might implies that using a verb as the target and using a noun as the target might be seen slightly different for the language model.

Problem 2

Report your **best setting** and its corresponding CIDEr & CLIPScore on the validation data. Briefly introduce your method. (TA will reproduce this result) (5%)

Setting	CIDEr & CLIPScore
Image encoder: "vit_large_patch14_clip_224.openai_ft_in12k_in1k", Lora: for each lora layer in the decoder ● rank=8 ● lora_attn_alpha=16, ● lora_dropout=0.2	CIDEr: 0.986 CLIPScore: 0.733

```python
class Attention(nn.Module):

    def __init__(self, cfg):
        super().__init__()
        self.c_attn = lora.MergedLinear(cfg.n_embd, 3 * cfg.n_embd,
                                        r=cfg.lora_attn_dim,
                                        lora_alpha=cfg.lora_attn_alpha,
                                        lora_dropout=cfg.lora_dropout ,
                                        enable_lora=[True, False, True])
```

```python
class Block(nn.Module):
    def __init__(self, cfg):
        super().__init__()
        self.ln_1 = nn.LayerNorm(cfg.n_embd)
        self.ln_2 = nn.LayerNorm(cfg.n_embd)
        self.attn = Attention(cfg)
        # multi-layer perceptron
        self.mlp = nn.Sequential(
            collections.OrderedDict(
                [
                    ("c_fc", lora.Linear(cfg.n_embd, 4 * cfg.n_embd, r=cfg.lora_attn_dim)),
                    ("act", nn.GELU(approximate="tanh")),
                    ("c_proj", lora.Linear(4 * cfg.n_embd, cfg.n_embd, r=cfg.lora_attn_dim)),
                ]
            )
        )
```

Model structure:

After input images are being encoded, they will be projected to text-embedding space by a projection layer (linear layer). The next step is to concatenate the image embedding with the text embedding, and it is important to note that the image embedding should be placed ahead of the text embedding in order to allow the decoder to learn the relationship between the image embedding and the targeted output text, ensuring that the decoder is trained to predict the next token.

Implementation detailed:

- Pretrained on projection layer only: before PEFT using lora, I pretrain the projection layer for 5 epoch
- Lora on Query and Value projection: I add lora layer on attention block(query and value projection matrices) and the linear mlp layer in Block module.
- Casual masking: in Attention module, I modify the original "self.register_buffer('bias, torch.tril(torch.ones(size, size)).view(1, 1, size, size))"
  to "self.register_buffer('mask', torch.tril(torch.ones(size, size)).view(1, 1, size, size))" in order to ensure the decoder's ability to caption the image by predicting the next possible text output based on the image embedding and the unmasked text embeddings.

Report **2 different attempts** of LoRA setting (e.g. initialization, alpha, rank...) and their corresponding CIDEr & CLIPScore. (5%, each setting for 2.5%)

Setting	CIDEr & CLIPScore
Pretrain projection layer: ● epoch: 5 ● Learning rate: 0.0002 PEFT: ● epoch: 5 ● lora rank=16 ● lora para lr: 0.0002 ● projection layer lr: 0.0001 ●	CIDEr: 1.036 CLIPScore: 0.726
Pretrain projection layer: ● epoch: 5 ● Learning rate: 0.0002 PEFT: ● epoch: 5 ● lora rank=8 ● lora_attn_alpha=16, ● lora_dropout=0.2 ● lora para lr: 0.0002 ● projection layer lr: 0.0001	<BEST> CIDEr: 0.986 CLIPScore: 0.733

Problem 3

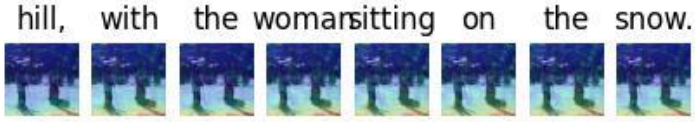
1. Given five test images ([p3_data/images/]), and please visualize the **predicted caption** and the corresponding series of **attention maps** in your report with the following template

The following five graph are the visualization of p1 model:


start A sheep and a lamb are


standing in a grassy field.


start A man and woman are standing on a snowy

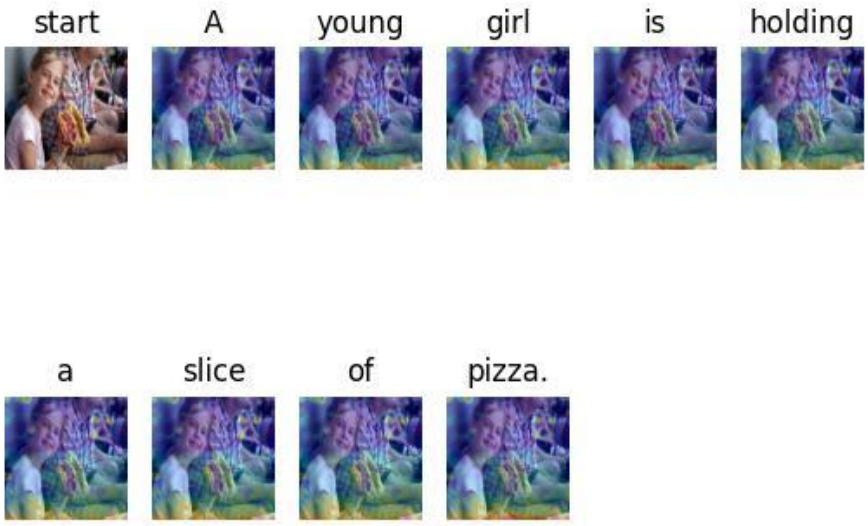

hill, with the woman sitting on the snow.


start A woman wearing a purple umbrella

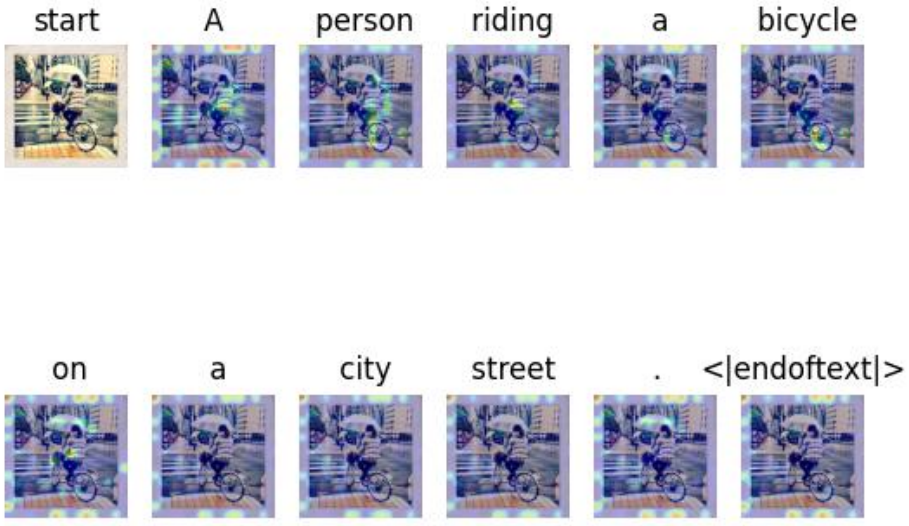

is walking down the street. snow.

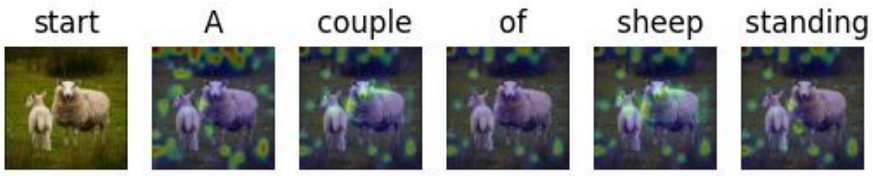

start A woman riding a


bike in the rain.

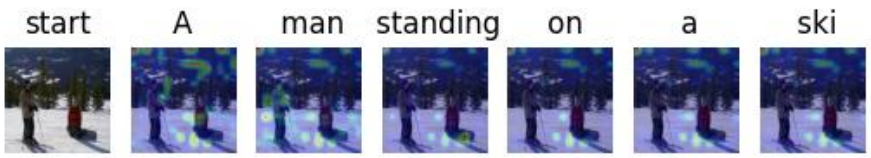


The following five graph are the visualization of p2 model:




start
A
couple
of
sheep
standing


next
to
each
other
.
<|endoftext|>


start
A
man
standing
on
a
ski


slope
with
a
snow
board
.
<|endoftext|>


start A woman with a hat and a

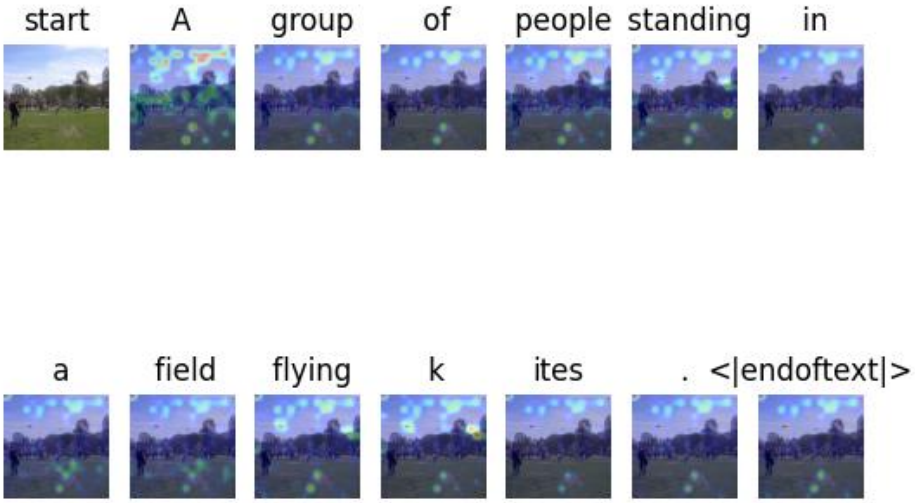

rain coat is holding an umbrella .<|endoftext|>


start A little girl and a


little boy eating pizza . <|endoftext|>

2. According to CLIPScore, you need to:
   - visualize top-1 and last-1 image-caption pairs
     - top-1: 000000001086.jpg
     - last-1: 000000000693.jpg
   - report its corresponding CLIPScore in the validation dataset of problem 2. (3%)
     - (1.0205078125, 'hw3_data/p2_data/images/val/000000001086.jpg')
     - (0.38665771484375, 'hw3_data/p2_data/images/val/000000000693.jpg')
3. 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? (3%)
   For the best top1 image, the result attention mask correspond to the image at the words “kite” and “flying”, but the rest of the words seems to be unrelated to the mask.
   On the other hand, the attention map of the image with the least score is less related to the predicted sentence.

Id : 1086


start A group of people standing in
a field flying k ites . <|endoftext|>

Id:693


start A woman with a pony tail and a
face that is black and white <|endoftext|>