DLCV HW3 Report

B11901040 項達均

November 19, 2024

1

1.1

LLAVA first encodes images with the pre-trained CLIP visual encoder ViT-L/14, which outputs visual features. The image features are then passed through a linear layer into the word embedding space. The image embedding and the language instructions are then inputted into the Vicuna LLM to generate the Language Response. The training of LLAVA is split into two stages:

1. **Pre-training for Feature Alignment.** The projection matrix is trained on a filtered CC3M dataset such that the image features can be aligned with the LLM word embedding.
2. **Fine-tuning End-to-End.** Both the pre-trained weights of the LLM and the projection layer are updated, while the visual encoder is frozen.

1.2

I experimented with two different prompts:

- **Write a description for the photo in one sentence.** This resulted in a CIDEr of 1.167 and ClipScore of 0.774.
- **Write a description for the photo.** This resulted in a CIDEr of 2.64×10^{-7} and a ClipScore of 0.798. This is because without telling the LLM to describe the photo in one sentence, the output becomes very long, spanning a few sentences. This led to a low consensus between the prediction caption and the ground truth, and thus its CIDEr is low. However, the ClipScore is higher than that of the other prompt, which suggests that longer outputs can describe the photo in more detail and correctness.

2

2.1

I used CLIP-Large with 16×16 patches as the vision encoder. The model is trained for 6 epochs with a learning rate of 5×10^{-3}. Greedy decoding strategy is used and a repetition penalty of 1.1 is used. Lora with rank 16 and alpha equal to 1 is added to the attention layers. The image projection layer is a two-layer perceptron trained from scratch. The training resulted in a CiDer of 0.997 and a ClipScore of 0.737.

2.2

I also experimented with $r = 2$ and $\alpha = 0.9$, this resulted in a CiDer of 0.989 and a ClipScore of 0.727. This suggests that the choice of r and α does not greatly impact performance, which agrees with the paper's results.

3

3.1

3.1.1 Problem 1

Figure 1: bike.jpg

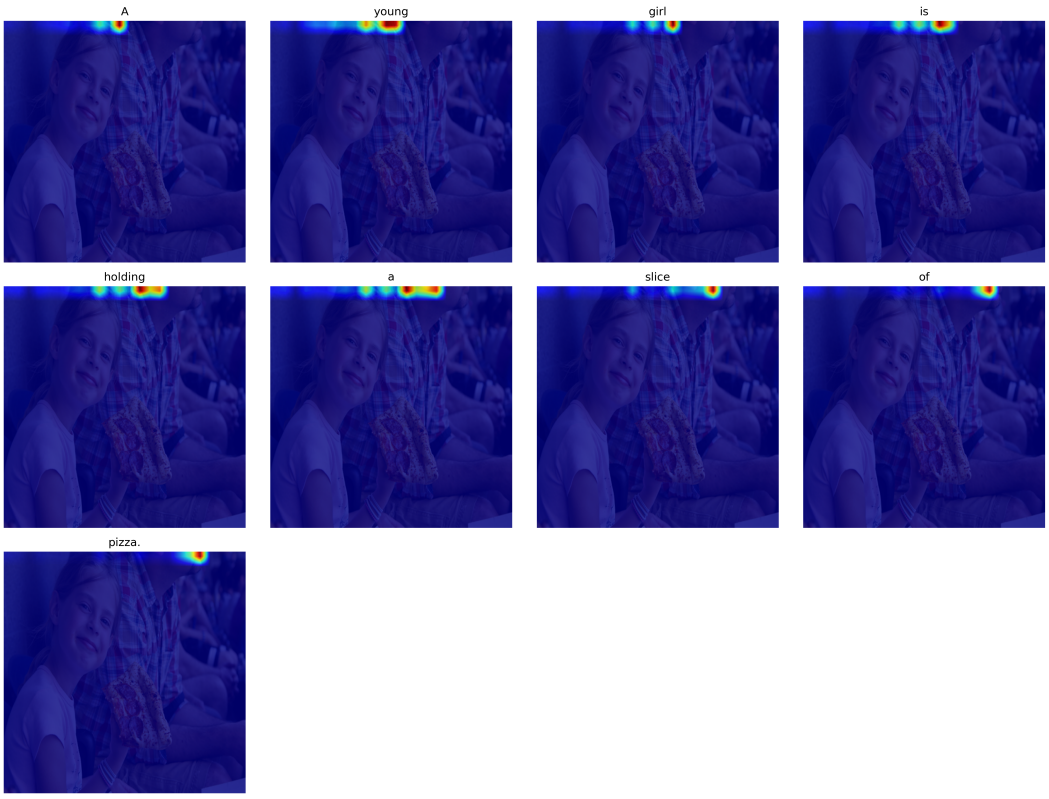

Figure 2: girl.jpg

Figure 3: sheep.jpg

Figure 4: ski.jpg

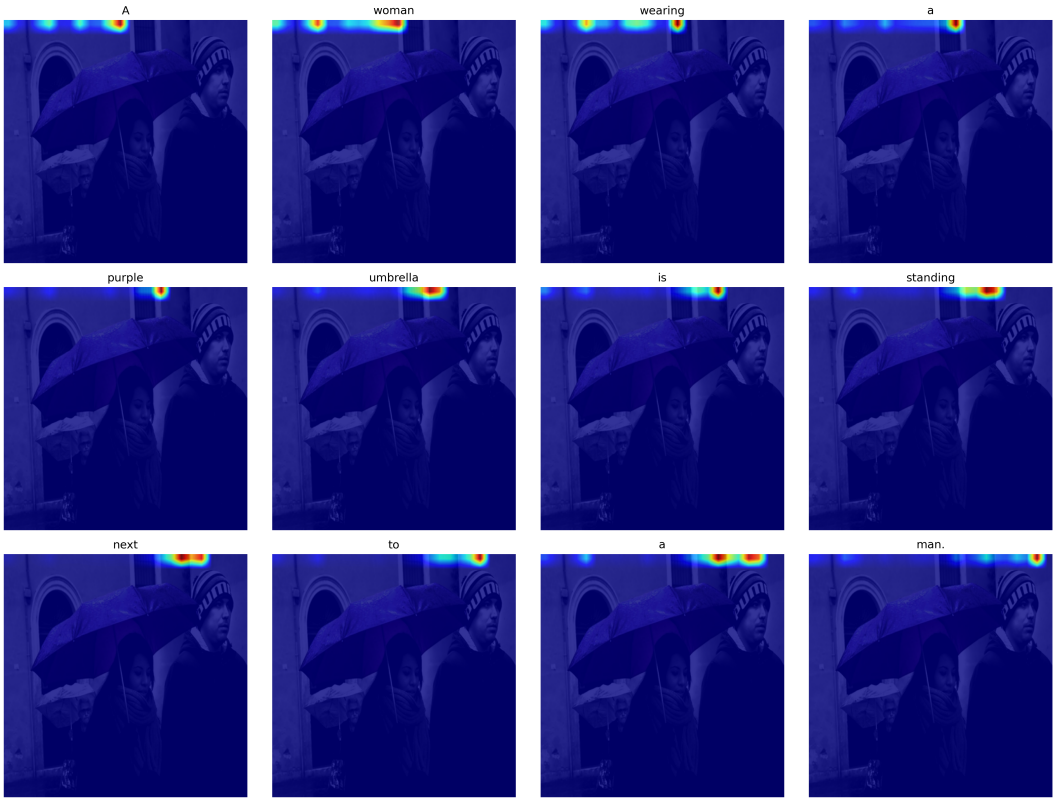

Figure 5: umbrella.jpg

3.1.2 Problem 2

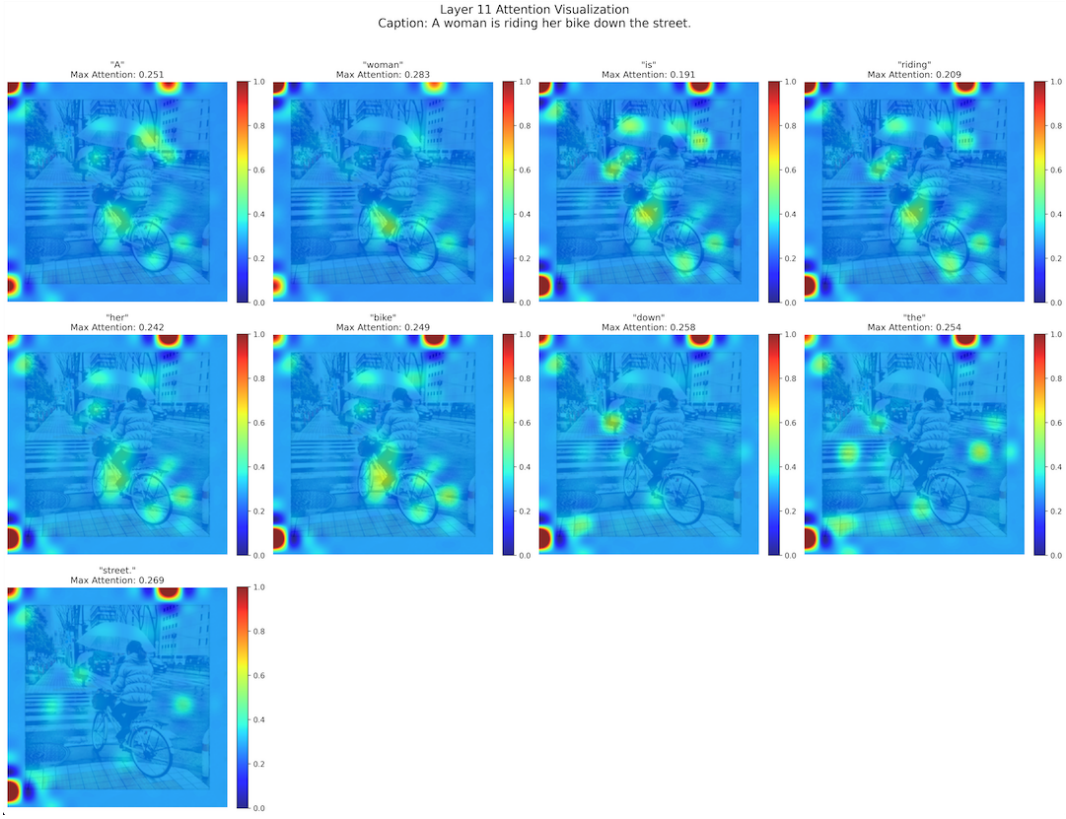

Figure 6: bike.jpg

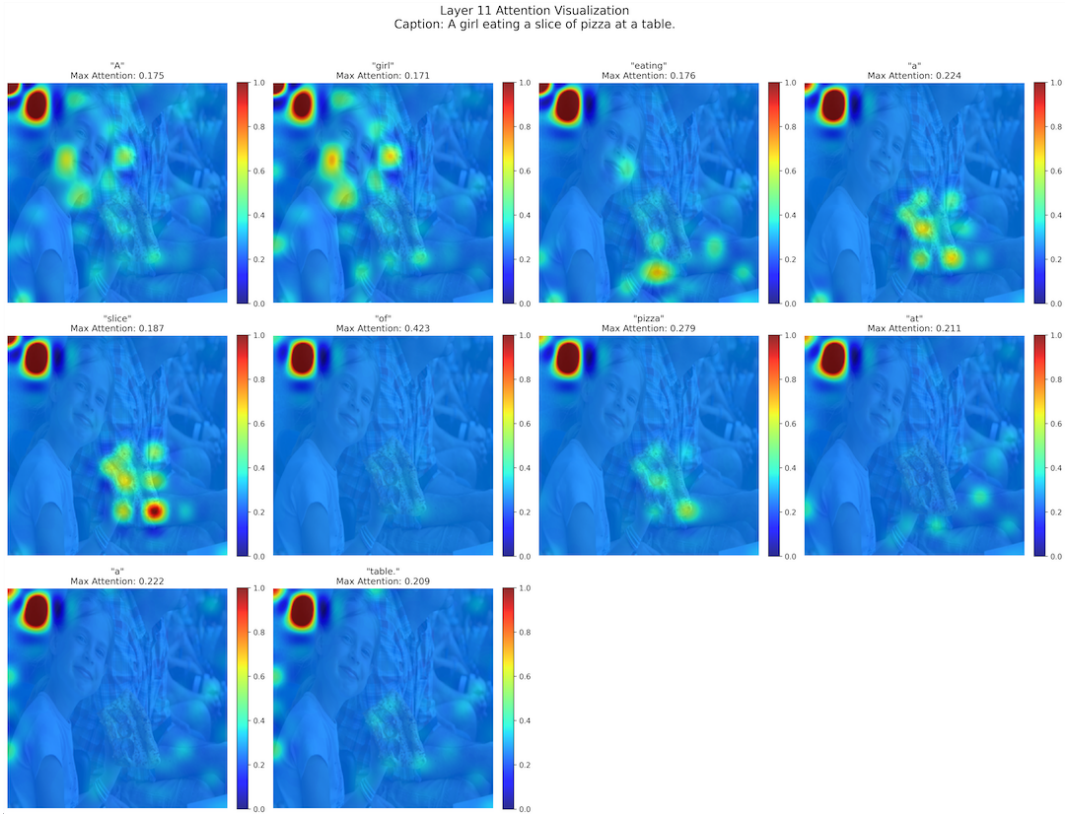

Figure 7: girl.jpg

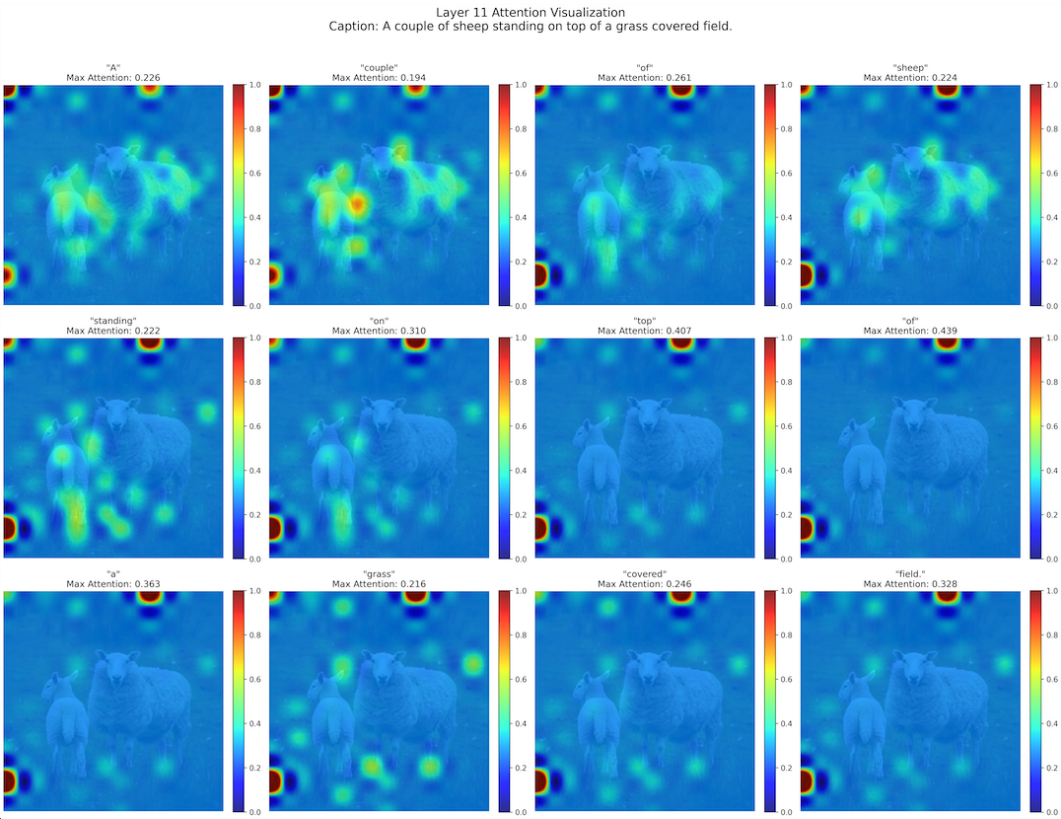

Figure 8: sheep.jpg

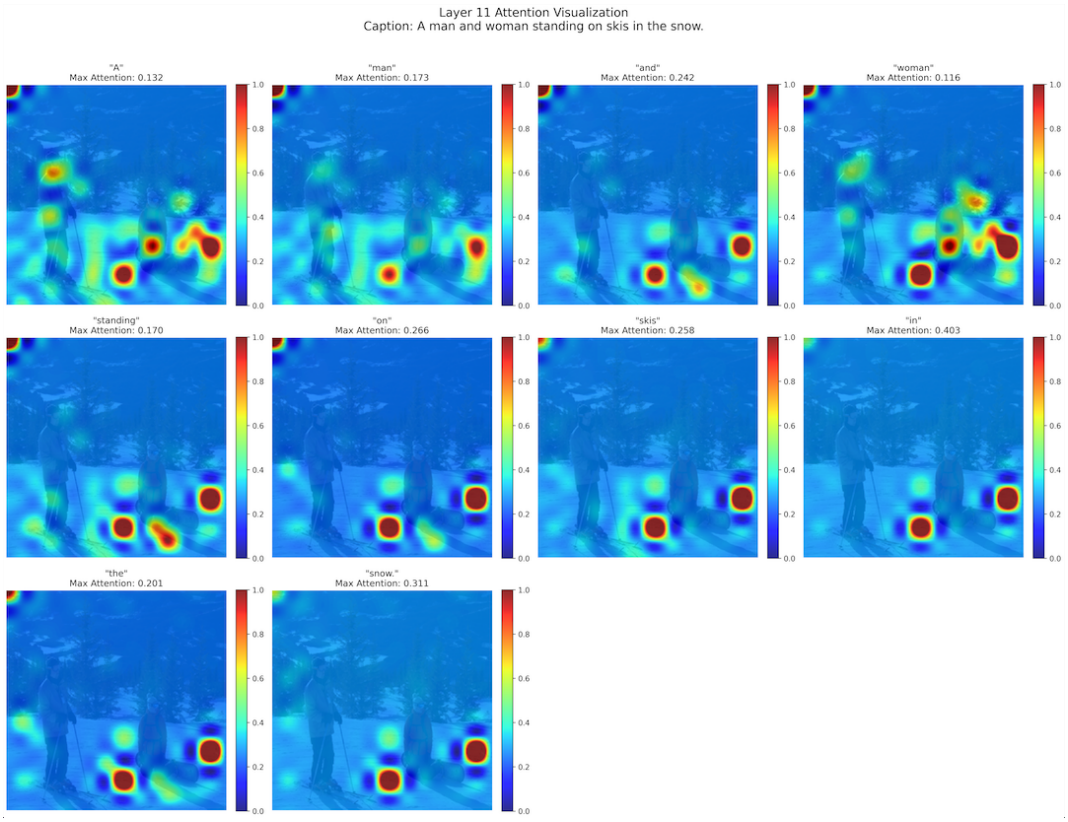

Figure 9: ski.jpg

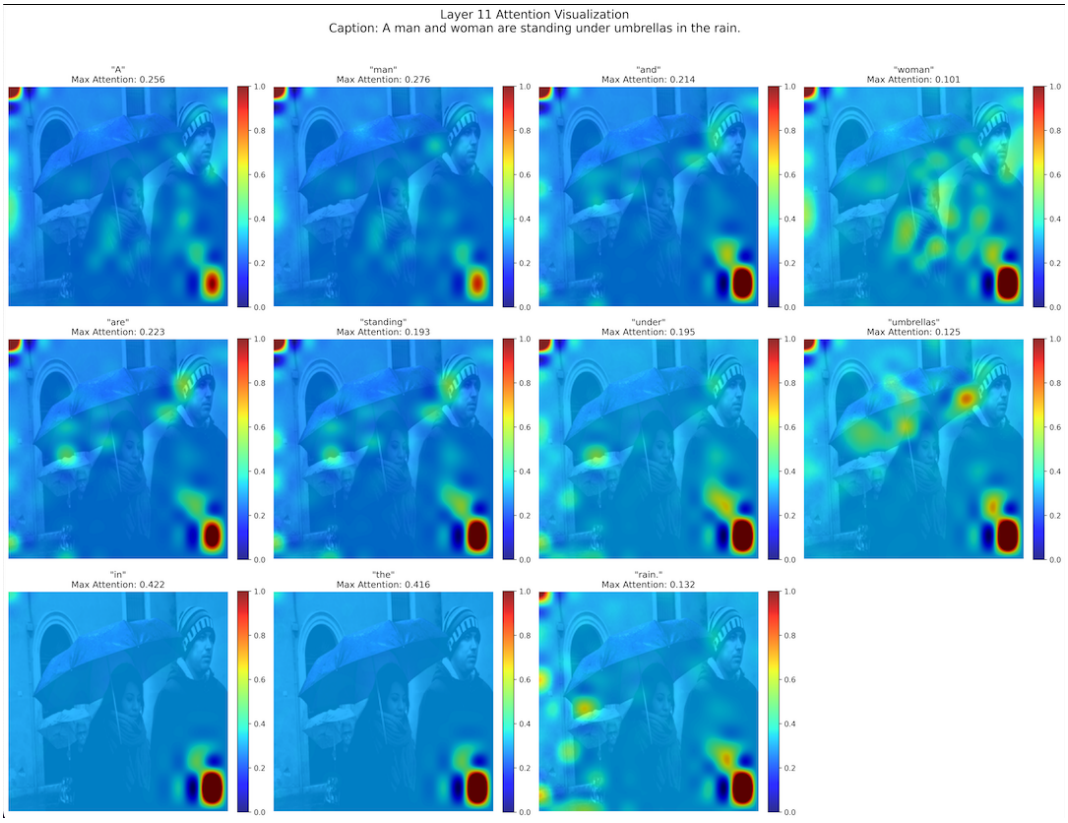

Figure 10: umbrella.jpg

3.2

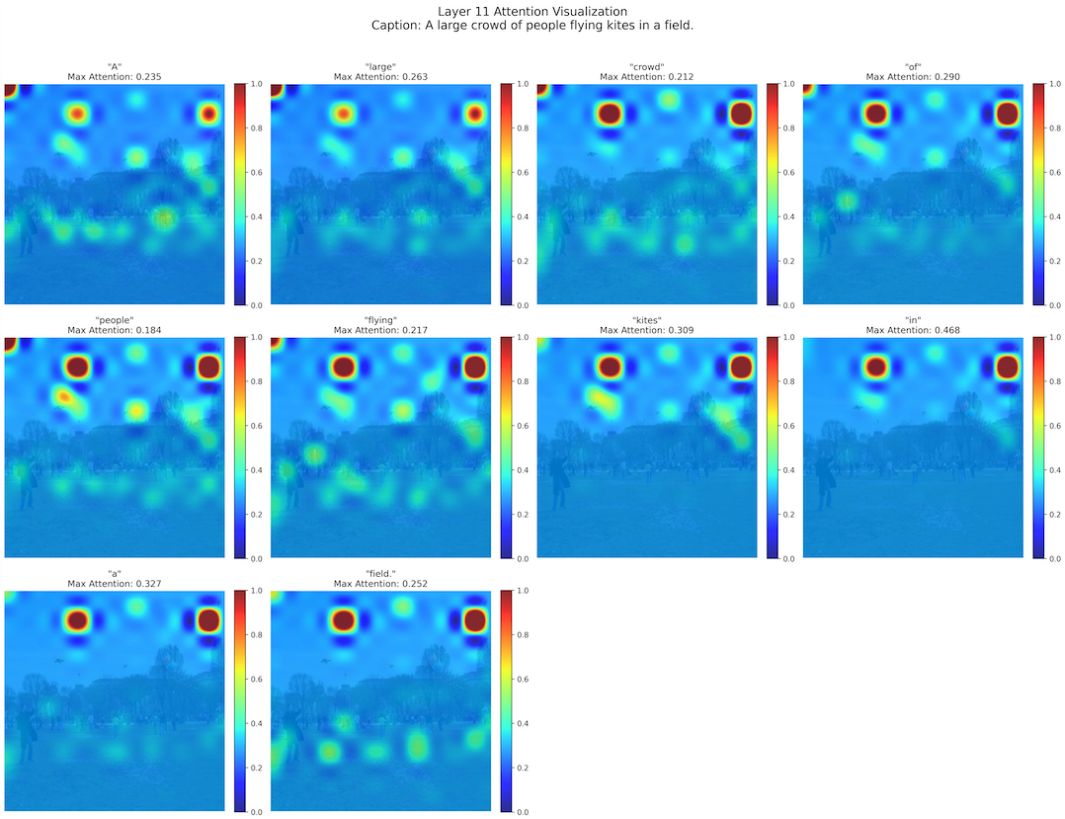

Figure 11: Best

The best image-caption pair had a CLIPScore of 1.005.

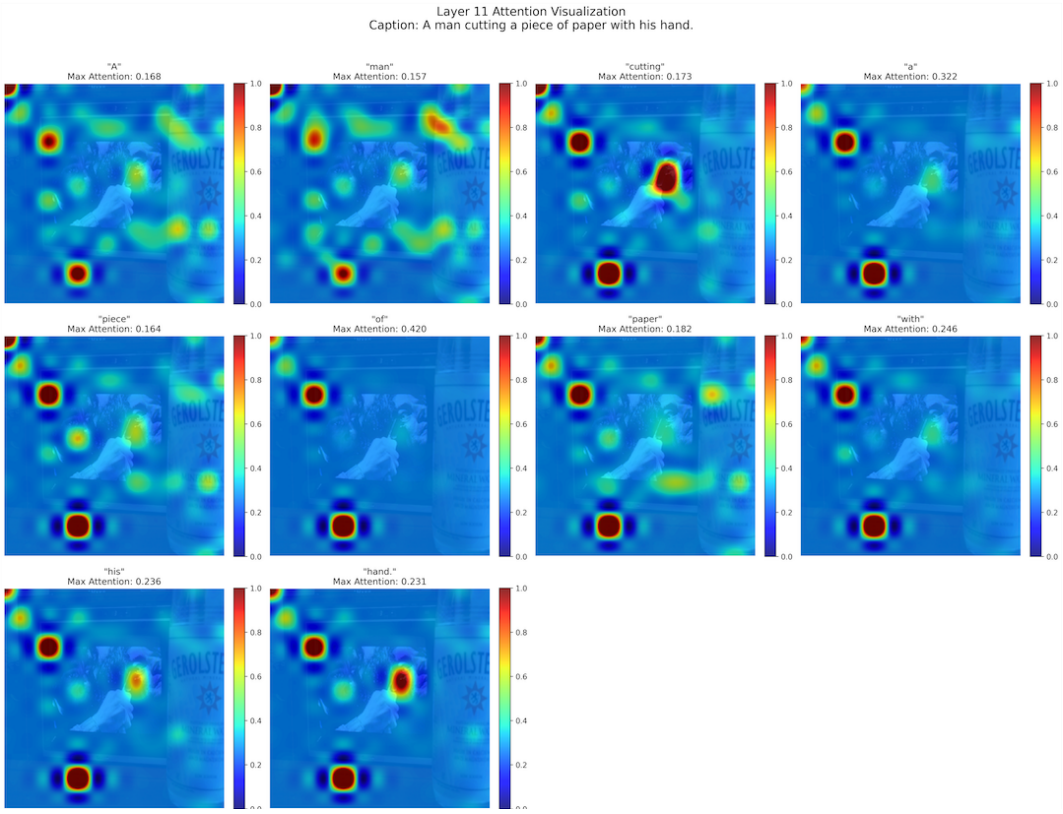

Figure 12: Worst

The worst image-caption pair had a CLIPScore of 0.401. In the image-caption pair with the highest CLIPScore, the attention map attends to the region corresponding to each word more accurately. For example, for the word "kites", the attention is more focused on the upper region of the image, corresponding to the region of the kites. And for the word "field", the attention is more focused on the lower region of the image, which is where the field is. However, in the image with the worst CLIPScore, the attention is more scattered for each word and doesn't attend to specific regions, meaning the model is unable to create a good caption that corresponds well to the image.