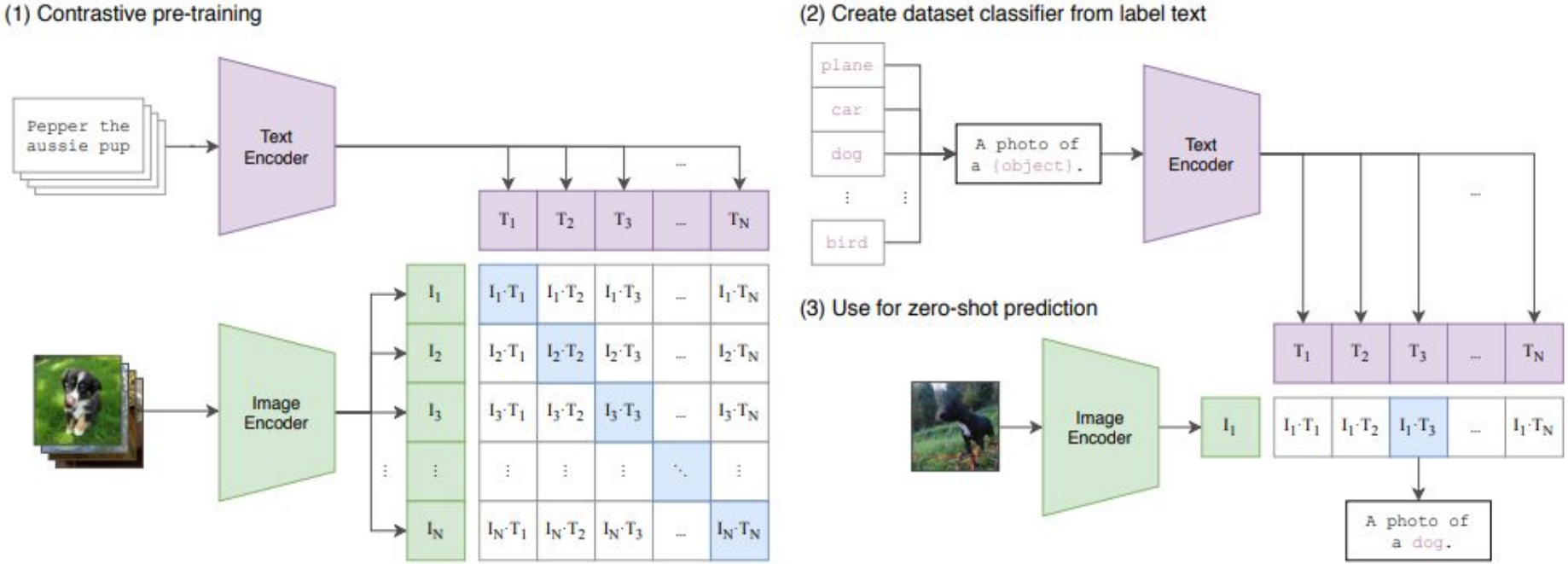
Evaluation: CLIP score and CLAP score

9/20/2024, Thanmaya, Mus2Vid

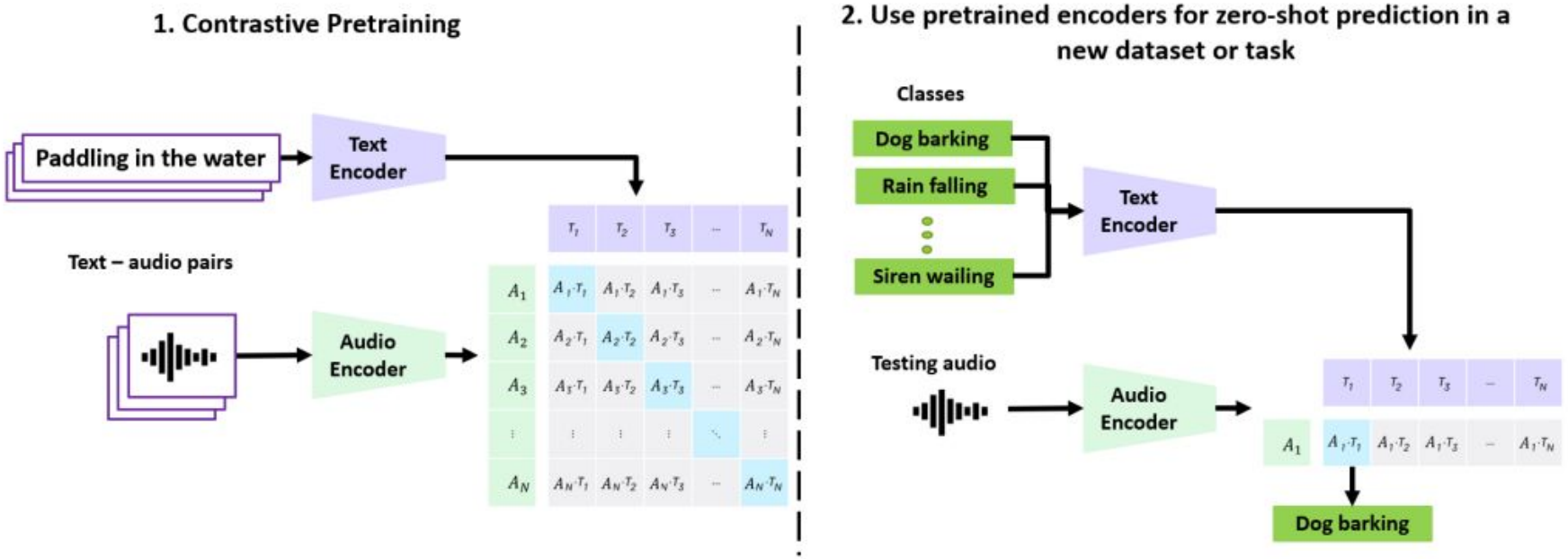
Overview of progress on tasks

- Looked into CLAP
- Worked through CLIP implementation to work on my machine
- Brainstormed ways to omit text from evaluation

CLIP vs CLAP score



CLIP vs CLAP score



How is similarity found in both?

- Cosine Similarity by comparing the vectors from the input (audio or image) to all of the class labels

$$\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

CLIP implementation

```python
def get_clip_similarity(image_path, text_input):
    # Load and process the image
    image = load_image(image_path)

    # Encode text input
    text = clip.tokenize([text_input]).to(device)

    # Get embeddings for both
    with torch.no_grad():
        image_embedding = model.encode_image(image)
        text_embedding = model.encode_text(text)

    # Compute cosine similarity between the embeddings
    similarity = torch.nn.functional.cosine_similarity(image_embedding, text_embedding)
    return similarity.item()
```

Note

- CLAP implementation follows similar structure But there are still bugs in my code

How could they be combined together to omit text encoder?

Utilize Pretrained Encoders:

- Use CLAP's audio encoder to embed audio inputs
- Use CLIP's image encoder to embed images

Leverage Learned Multimodal Spaces:

- Both CLIP and CLAP have learned to map their respective modalities (image/audio) to a space that aligns with text embeddings
- These spaces may capture information without explicitly needing the text

Next Steps:

- Get CLAP to work on my machine
- Test CLIP on data that exists within our project
- Any recommendations?