Evaluation: Implementation

10/18/2024, Thanmaya, Mus2Vid

Problem statement

- Find a metric to see how similar image and audio pairs are
- In order to see how our model is performing

Progress

Last week's Issues:

- IMSM still produced odd issues where scores were not as expected for MelBench Data set

This week's progress:

- Found CLIP and CLAP scores for our model output
- Preprocessed our model output

Implementation Details

- Takes in a folder of images (for 1 video there were about 74 generated images)
- Takes in an audio file, and list of text inputs (from the prompts)
- Outputs CLIP and CLAP similarity between each line of text to each image
    - Assumptions: prompts generate images in order

Tasks yet to complete:

- Normalize scores
-

Examples

Low CLAP score: -4.809

- “A man is sitting at a desk, writing a letter. He is wearing a suit and tie and looks very serious.”

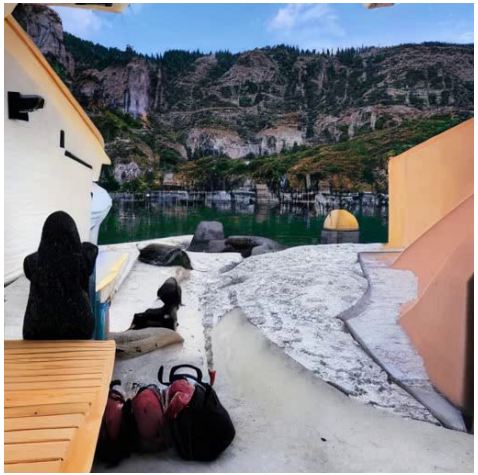
High CLAP score: 3.364

- “Slow, classical, violin, cello, flute, oboe, sadness, heartbreak, love, man, woman”

Examples

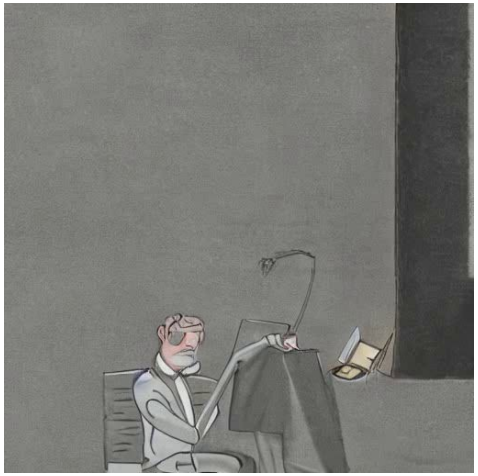
Low CLIP score:

- “Slow, calm, peaceful, relaxed, meditative, classical, slow-paced”



High CLIP score:

- “A man is sitting at a desk, writing a letter. He is wearing a suit and tie and looks very serious. B”



Next Week

- Normalize CLIP and CLAP scores
- Gather average score throughout the video
- Pipeline is little complicated (preprocessing occurs separately from score calculation)