1. VIRAT (https://viratdata.org/)
    a. Videos of pedestrians in public spaces
    b. Videos are annotated for
        i. Scene elements (sky, building, shadows, grass, etc)
        ii. Object types (door, tree, parking meter, etc)
        iii. Activities of people (standing, crouching, sitting, etc)
        iv. Activities of vehicles (moving, stopping, starting, etc)
    c. 8.5 hours of video in 11 different outdoor scenes annotated for 12 event types

2. Extended Cohn-Kanade dataset (https://www.kaggle.com/competitions/visum-facial-expression-analysis/data)
    a. 593 video sequences of 123 different people
    b. The video sequences start with a neutral expression and reach a peak targeted expression
    c. 327 of 593 are labeled with one of 7 expression classes (anger, fear, happiness, etc.)

3. Weizmann dataset (https://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html)
    a. 90 video sequences of 9 different people performing 10 actions such as running, skipping, bending, etc.
    b. Each video sequence consists of one person performing one action

4. Multi-model Intent Understanding dataset (https://github.com/apple/vqg-multimodal-assistant/blob/main/data/apple/apple_dev_all_keyword.csv)
    a. 12,000 images with 44,000 natural language questions related to the images
    b. Each question is annotated based on the underlying intent of the question (for example does the person want factoid/descriptive information, are they searching for a local business, asking for the recipe of a food item, etc.). There are 14 possible intent categories