1. ImageNet Vid (link)
    a. 30 object categories in videos like aeroplane, antelope, bear, etc.
    b. The task is to annotate objects (there may be multiple) using bounding boxes for each frame

2. ImageNet - VidVRD (link)
    a. A subset of 1K videos from the ImageNet Vid dataset
    b. The dataset focuses on visual relations between objects in videos eg. person-touch-dog, cat-above-sofa
    c. The relations are represented as subject-predicate-object
    d. 35 categories of subjects/objects and 132 categories of predicates

3. Youtube - Objects dataset (link)
    a. Dataset of Youtube videos with the objects in them weakly annotated
    b. There are 10 object classes with 9 - 24 videos for each class (eg. bird, boat, car). 155 videos in total
    c. Weak annotations mean that only the class of the object in the video is specified for a video. In addition to this they include all candidate bounding boxes and the one chosen by the method the authors proposed

4. OAK dataset (link)
    a. First person, outdoor video dataset with bounding box annotations for 80 video snippets (17.5 hours)
    b. 105 object categories

5. VisDrone dataset (link)
    a. Large scale UAV video dataset. More complex and crowded than ImageNet Vid
    b. Has 288 video clips which are annotated with bounding boxes for common objects like pedestrians, vehicles, bicycles, etc.

6. NYU Depth Dataset v2 (link)
    a. RGBD videos with image segmentation annotations
    b. It contains video sequences of a variety of indoor scenes
    c. 1,449 labelled RGBD video sequences

7. YFCC100M (link)
    a. 100 million images and 800K videos along with descriptions
    b. Not a high quality dataset; descriptions are not reliable

8. MOT dataset (link)
    a. Multi-object tracking dataset with 22 video sequences
    b. Annotations are available only for the 11 training sequences

9. ICDAR2013 Video (link)
    a. Video text detection dataset. It is used for near-horizontal text detection
    b. It has 299 training images and 233 test images with word level annotations

10. MSRA-TD500 (link)
    a. Text detection dataset with 500 images (300 train + 200 test)

11. Flickr30k Entities (link)
    a. 158K captions, 244K co-references of the same object in an image and 276k manually annotated bounding boxes for mentions of objects
    b. The dataset can be used for finding natural language mentions of an object in an image

12. RefCOCO, RefCOCO+ and RefCOCOg (link)
    a. These datasets are built on the Microsoft COCO image dataset. Expression referring to objects in the image have been added to images in the COCO dataset
    b. Images with multiple objects of the same kind are selected so that expressions can refer to one of them
    c. RefCOCOg: 85K referring expressions for 54K objects in 26K images. Images have 2 to 4 objects of the same kind
    d. RefCOCO+: RefCOCO+ has referring expressions without any location words so the referring expressions describe objects purely based on their appearance. It has 141K expressions for 49K objects in 19K images
    e. RefCOCO: It has 142K expressions for 50K objects in 19K images

13. Google Referring Expression dataset (link)
    a. Referring expression dataset that builds on the Microsoft COCO dataset with 123K images

14. ReferIt dataset (many papers reference it but I couldn’t find it) (link to paper)
    a. Has images with referring expressions for objects in them

15. Visual Genome (link)
    a. Image dataset with bounding box annotations for objects and actions
    b. 100K images with 4 million natural language object descriptions for 1.3 million instances of 75K objects
    c. It also has natural language descriptions of relations between objects in images